REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 2000	3. REPORT TYPE AND DATES COVERED Conference Proceeding

4. TITLE AND SUBTITLE

Paradigms of Complexity: Fractals and Structures in the Sciences

5. FUNDING NUMBERS

N00014-00-1-1003

6. AUTHOR(S)

Miroslav M. Novak, Editor

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

World Scientific Publishing Co., Pte. Ltd.
Suite 1B, 1060 Main Street
River Edge, NJ 07661

8. PERFORMING ORGANIZATION REPORT NUMBER

ISBN 981-02-4292-1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research International Field Office
PSC 802 Box 39
FPO AE 09499-0039

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited.
Government Purpose Rights License. All other rights reserved by the copyright holder World Scientific Publishing Co., Pte. Ltd.

12b. DISTRIBUTION CODE

A

12. ABSTRACT (Maximum 200 words)

This document contains all accepted papers to Fractal 2000, the 6th International Multidisciplinary Conference, held 16-19 April 2000 in Singapore. One of the aims of the conference was to review the current status in the field of fractals within realm of complexity and to explore future directions. There is a plentitude of examples in nature displaying complex patterns. The couplings responsible for these intricate patterns give rise to processes and phenomena that are absent when couplings are not present. The universality of such systems can be observed over a wide range of scales, from sub-atomic domain to that of cosmology. Furthermore, in addition to physics, mathematics, and chemistry, where these phenomena were observed first, the existence of complex systems has been confirmed in many disciplines such as medicine, biology, economics and sociology. The underlying geometry is frequently non-Euclidean and can be best analyzed using the tools of fractal analysis.

14. SUBJECT TERMS

Fractals, Mathematics, Scientific analysis

15. NUMBER OF PAGES
307

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19, SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Paradigms
of Complexity

Fractals and Structures
in the Sciences

Editor

Miroslav M. Novak
School of Mathematics, Kingston University, UK

20010501 138

World Scientific
Singapore • New Jersey • London • Hong Kong

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Paradigms of Complexity: Fractals and Structures in the Sciences

To Doreen, Karl and Erika

Contents

THE ORIGIN OF COMPLEXITY

LEON O. CHUA

Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley,
Berkeley, CA 94720, USA
E-mail: chua@eecs.berkeley.edu

Nature abounds with complex *patterns* and *structures* emerging from *homogeneous media* operating far from thermodynamic equilibrium. Such phenomena, which are widely observed in both inanimate (non-biological) and biological media, can be modeled and studied via the CNN (Cellular Neural/Nonlinear Network) paradigm in an in-depth and unified way. Whether a homogeneous medium is capable of exhibiting complexity depends on whether the CNN *cells*, or its *couplings*, is *locally active* in a precise *mathematical* sense. This *local activity principle* is of universal generality and is responsible for all symmetry breaking phenomena observed in a great variety of non-equilibrium media ranging from the emergence of negative differential conductance in bulk semiconductor materials (e.g., Gallium Arsenide in Gunn Diodes) to the emergence of artificial life itself. The main result of this paper consists of a set of explicit *analytical conditions* for calculating the *parameter ranges necessary for the emergence* of a non-homogeneous static or dynamic pattern in a homogeneous medium operating under an influx of energy and/or matter. The resulting "complexity related" inequalities are applicable to all media, continuous or discrete, which have been mapped into a CNN paradigm.

> *One of the most interesting aspects of the world is that it can be considered to be made up of patterns. A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made rather than by the intrinsic nature of these elements.*
>
> *Norbert Wiener*

1 Introduction

How does the leopard get its spots? How does the zebra get its stripes? How does the fingerprint get its patterns? How does an ant colony manage to *self-organize* into an impressive pattern of activities when individual ants are known to be quite dumb? How does an initially mixed distribution of black and white population in a housing community manage to redistribute over time into segregated black and white neighborhoods with well-defined boundaries?

How does a bee sting at a finger tip trigger the propagation of an electrical impulse to the brain of a healthy person, and how does this distress signal fail to propagate in the nerves of patients suffering from multiple sclerosis? How do some members of a colony of starving amoeba send out a target and spiral wave signal to attract neighboring amoebae and transform them into spores, and then to regenerate into amoeba again when food (bacteria) becomes available? How does a "scroll wave" get generated in the cardiac muscle by the inadvertent presence of an electrical impulse during a *vulnerable window* of a few milliseconds, often leading to sudden cardiac death?

The above phenomena are some manifestations of a multidisciplinary paradigm called *emergence*, or *complexity*. They share a common unifying principle characteristic of *dynamic arrays*, such as *cellular neural networks*, namely, interconnections of a sufficiently large numbers of *simple* dynamical units, which can exhibit extremely complex, synergetic, and *self-organizing* behaviors. The common denominator in all of these pattern formation and active wave propagation phenomena is the presence of an *active medium*, powered by a constant supply of external energy. For example, in the brain, the active medium is provided by a sheet-like array of massively interconnected excitable neurons whose energy comes from the burning of glucose with oxygen. In cellular neural networks, the active medium is provided by the local interconnections of active cells, whose building blocks include *active* nonlinear devices (e.g., CMOS transistors) powered by batteries.

Research on *Emergence* and *Complexity* has gained immense momentum during the past decade[1−15]. The fundamental problem is to uncover nature's secret mechanisms which are responsible for the self organization and spontaneous emergence of many stable complex (*static* and *dynamic*) patterns[a] in homogeneous media operating *far-from-thermodynamic equilibrium* [16]. Indeed, nature is abound with all sorts of patterns ranging from regular snow flakes to chaotic brain waves[17,18,19]. Understanding and controlling such patterns is essential for designing new generations of brain-like molecular devices and systems endowed with artificial intelligence and self-repair capabilities.

The *homogeneous* media alluded to above consists usually of an active *bulk* medium (e.g., bulk materials with *negative* resistivity, such as Gallium Arsenide in Gunn Diodes[20], nerve membranes, heart tissue layers, chemical mixtures in stirred reactor tanks, etc.) modeled by one or more nonlinear *PDE* 's (partial differential equations) where the spatial coordinate, as well as the state variables, are represented by *continuous real numbers*. What is truly fascinating is that while these active media are completely unrelated—they can range from inanimate materials to living tissues[21]—yet the patterns they exhibit tend to resemble each other under appropriate initial and boundary conditions. It makes sense therefore to hypothesize that a common mechanism must be responsible for the emergence of each type of patterns (e.g., Turing patterns, spiral waves, etc.)[b]. This remarkable observation motivates the development of a *unified paradigm* capable of exhibiting most, if not all, static and dynamic patterns (i.e., *dissipative structures*) in active homogeneous media operating far from thermodynamic equilibrium. Such a paradigm has been developed recently and is the subject of a recent treatise [22]. The paradigm is dubbed the *CNN*, an acronym for *cellular neural networks* when used in the context of brain science, or *cellular nonlinear networks*, when used in other more general contexts.

A CNN is defined by *two* mathematical constructs:

1. A spatially discrete collection of nonlinear dynamical systems called *cells*, where *information* can be encrypted into each cell via 3 independent variables

---

[a]Such patterns are called *dissipative structures* by Ilya Prigogine[14] because *energy must be continually supplied and dissipated* in order to maintain such structures.

[b]For related works on emergence and complexity from different perspectives, the reader is referred to the very readable expositions by Crick[8], Eigen[10], Gell-Mann[7], and Prigogine[14].

Sphere of influence S_{ijk} of cell C_{ijk} containing all cells $C_{\alpha\beta\gamma} \in S_{ijk}$
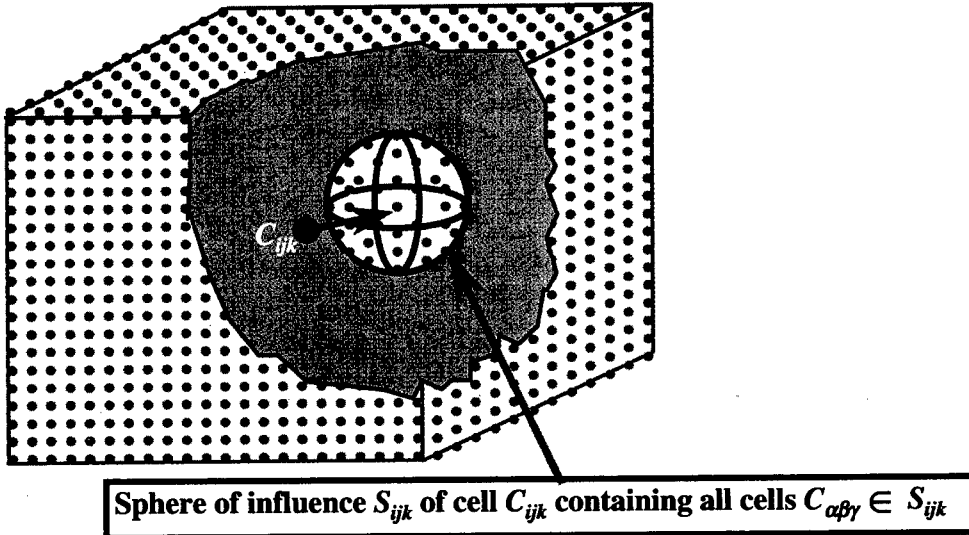
Figure 1. A CNN with a cutout view which exposes an inner cell C_{ijk} and its sphere of influence S_{ijk}. Only those cells $C_{\alpha\beta\gamma}$ located within S_{ijk} are coupled to cell C_{ijk}.

called *input, threshold,* and *initial state.*

2. An *interconnection law* relating one or more relevant variables of each cell C_{ij} to all neighbor cells C_{kl} located within a prescribed sphere of influence $S_{ij}(r)$ of radius r, centered at C_{ij}.

In the special case where the CNN consists of a homogeneous array, and where its cells have no inputs, no thresholds, and no outputs, and where the sphere of influence extends only to the *nearest* neighbors (i.e., $r = 1$), the CNN reduces to the familiar concept of a *lattice dynamical system* from mathematics.

The schematic diagram of a 3-dimensional CNN is shown in Fig.1, where a typical *cell C_{ijk}* is highlighted along with its *sphere of influence S_{ijk}*. Let us consider some examples.

Example 1. Emergence of 3-Dimensional Knot Patterns

Suppose each cell in Fig.1 consists of a Chua's circuit with three external terminals (one of them being the *ground* reference terminal), as shown in Fig.2. Since each cell can interact with its neighbors only through the 2 ungrounded terminals, each one coupled to a corresponding node of a neighbor cell via a resistor, each ungrounded terminal serves as a *port* where energy can flow into or out of the cell. Consequently, the 3-terminal circuit cell in Fig.3 is also called a 2-port in electrical engineering. Suppose each cell in Fig.2 is *coupled* to its 6 nearest neighbor cells (two along each coordinate axis) via positive linear resistances, as depicted in Fig.2.

4

A section of the coupling circuit

C(ijk)

Chua's Circuit

Sphere of influence S_{ijk} of cell *C(ijk)* containing all cells $C(\alpha\beta\gamma) \in S_{ijk}$
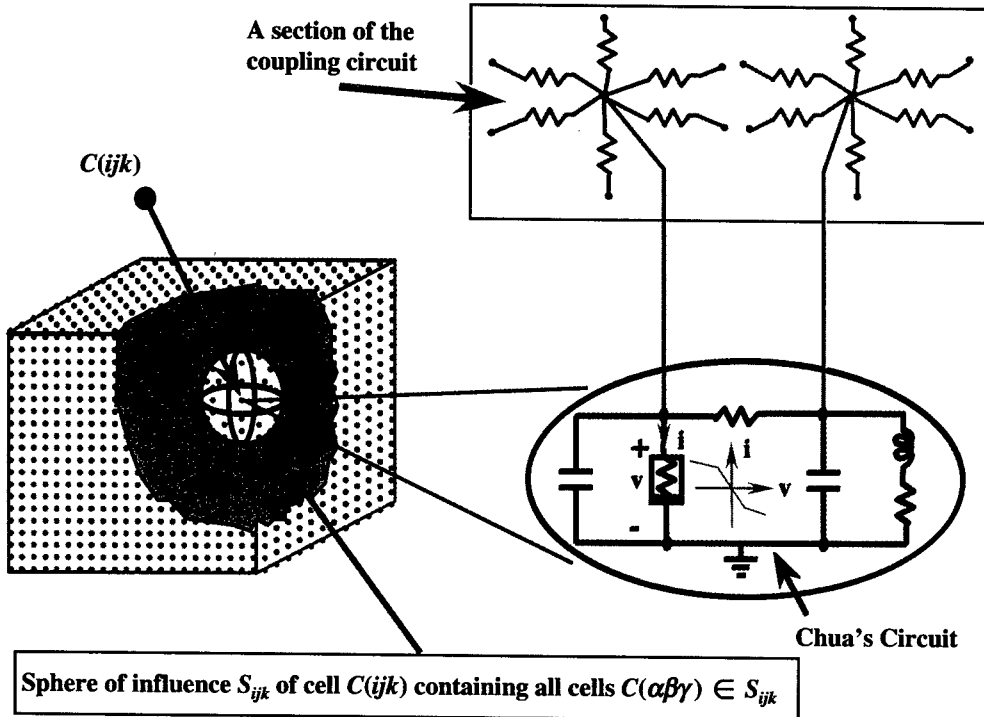
Figure 2. Each cell in this 3-dimensional CNN is a Chua's oscillator with 3 external terminals, or 2 ports, with a common ground. Each external terminal (except the ground), or port, of each cell is connected to a capacitor inside the cell, and *coupled* to a corresponding terminal of the 6 neighboring cells via 6 linear passive resistors.

If all elements inside the cell shown in the inset are passive, and if the nonlinear resistor is *not locally active* (e.g., a pn junction diode) in the sense of *Definition 1* in Sec.4.2, then it can be shown that given any *initial* voltage distributions, all node-to-datum voltages must tend to zero as $t \longrightarrow \infty$. In other words, this CNN must have a *homogeneous* (uniform) solution at all nodes. This homogeneity is generally expected since all cells in Fig.7, and their couplings, as well as boundary conditions, are identical, and there are *no input sources.*[c]

---

[c]We assume throughout this paper that the CNN has no input sources and has a symmetrical boundary condition (e.g., a zero-flux or Neumann boundary condition), since the central concept of *emergence* implies that any non-homogeneous output pattern must emerge via *self organization*, and not through any external input, or non-symmetrical boundary conditions.

(a)

(b)

(c)

(d)
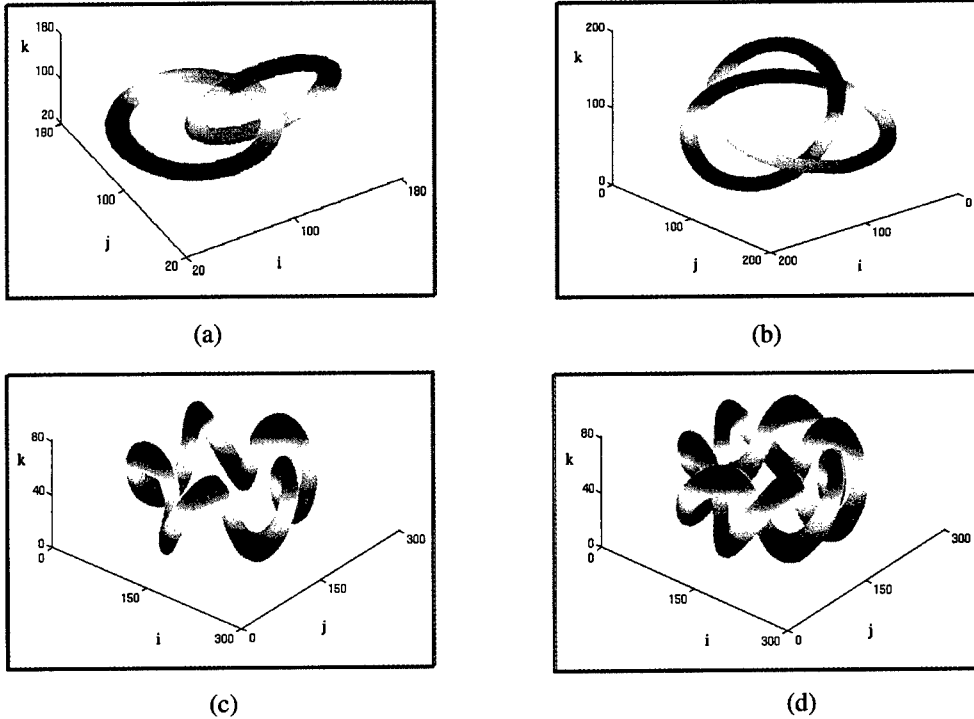
Figure 3. Three-dimensional *knot* voltage distribution patterns from a CNN cube made of Chua's oscillator cells, coupled via a linear passive resistive grid. (a) A Hopf-link pattern. (b) A 3-knot pattern. (c) A simple helix. (d) A double helix.

However, if the nonlinear resistor is chosen to be a Chua's diode [23], as shown in the inset in Fig.2, then for the *fixed* choice of circuit parameters [22], the 4 distinctly different 3-dimensional structures in Fig.3 can be obtained in steady state by choosing the 4 different sets of initial conditions [22]. Note that this CNN is symmetrical with respect to the center of the CNN cube in Fig.2 and there are no inputs. Yet we have a non-uniform constant steady state node-to-datum voltage distribution ! In the parlance of *complexity theory*, we say the homogeneous CNN undergoes a *symmetry breaking* and any *non-homogeneous* node-to-datum steady state voltage distribution is called a *pattern*, or a *dissipative structure* if the medium is *non-conservative* in the sense that *energy dissipation* is essential to maintain the structure.

(a)   (b)
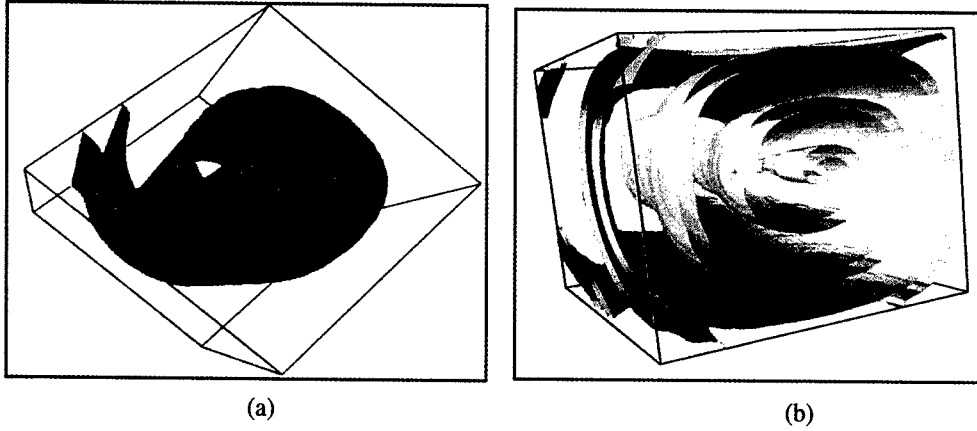
Figure 4. (a) A straight scroll vortex. (b) A twisted scroll wave.

Example 2. Emergence of 3-Dimensional Scroll Waves

Consider the same CNN cube as in Fig.2 except that in this case, only the port voltage across the nonlinear resistor(Chua's diode) is coupled to corresponding node voltages of the 6 neighboring cells. This is equivalent to *open circuiting* the second port (on the right) of the Chua's oscillator in Fig.2, so that each cell reduces to a *one-port* circuit with 2 external terminals (including the ground reference terminal). Using the *circuit parameters*, and the *initial* and *boundary* conditions [22], we obtain the two *scroll wave* structures shown in Figs.4(a) and 4(b), respectively, for one instant of time. Unlike the *stationary* structures shown in Fig.3, the scroll waves in Fig.4 represent an *active nonlinear wave* which evolves continuously with a scrolling structure at all times.

Example 3. Emergence of 2-Dimensional Spiral Waves

If we consider a 2-dimensional version of the preceding CNN cube, we would obtain the simplified 2-dimensional CNN shown in Fig.5, where each "Chua's oscillator"-one-port is represented by a 2-terminal black box, coupled to its neighbors by 4 linear positive resistances, two along each coordinate axis. Using the circuit parameters and the initial and boundary conditions given in [22], we obtain the *spiral wave* structure shown in Fig.6. Again, this is a *dynamic pattern* which rotates continuously for all times.
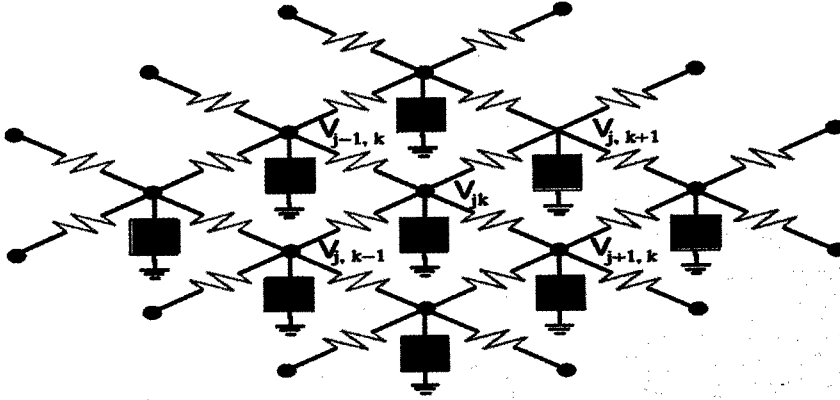
Figure 5. A two-dimensional CNN made of Chua's oscillator cells coupled via a single layer of linear resistive grid.

2 Mapping PDE Into CNN

A vast majority of active homogeneous media which are known to exhibit complexity in the form of *dissipative structures* are modeled by a *reaction diffusion PDE* [14-18]:

$$\frac{\partial x_i}{\partial t} = f_i(x_1, x_2, ..., x_n) + D_i \left(\frac{\partial^2 x_i}{\partial x^2} + \frac{\partial^2 x_i}{\partial y^2} + \frac{\partial^2 x_i}{\partial z^2} \right),$$

$$i = 1, 2, ..., n \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ are state variables, (x, y, z) are spatial coordinates, $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_n(\mathbf{x}))$ is a nonlinear vector function of $\mathbf{x}$ called the *kinetic term*, and $D_1, D_2, ..., D_n$ are constants called *diffusion coefficients*. Replacing the *Laplacian* in Eq.(1) by its *discretized* version

$$\frac{\partial^2 x_i}{\partial x^2} + \frac{\partial^2 x_i}{\partial y^2} + \frac{\partial^2 x_i}{\partial z^2} \longrightarrow (\nabla^2 \mathbf{x}_{\alpha,\beta,\gamma})_i \tag{2}$$

8

where

$$\nabla^2(\mathbf{x}_{\alpha,\beta,\gamma})_i \stackrel{\Delta}{=} x_i(\alpha+1,\beta,\gamma) + x_i(\alpha-1,\beta,\gamma) + x_i(\alpha,\beta+1,\gamma) + x_i(\alpha,\beta-1,\gamma) +$$
$$x_i(\alpha,\beta,\gamma+1) + x_i(\alpha,\beta,\gamma-1) - 6x_i(\alpha,\beta,\gamma)$$
$$i = 1,2,...,n \tag{3}$$

is the discrete *Laplacian operator* on the *ith* component x_i of the state variable $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ about the grid point with spatial coordinate (α,β,γ). We obtain the following associated *Reaction Diffusion CNN* equation

$$\dot{x}_i(\alpha,\beta,\gamma) = f_i(x_1(\alpha,\beta,\gamma), x_2(\alpha,\beta,\gamma), ..., x_n(\alpha,\beta,\gamma)) + D_i\nabla^2(\mathbf{x}_{\alpha,\beta,\gamma})_i \quad (4)$$

where $i = 1,2,...,n; \alpha = (1,2,...,N_\alpha), \beta = (1,2,...,N_\beta)$, and $\gamma = (1,2,...,N_\gamma)$.

Here $x_i(\alpha,\beta,\gamma)$ denotes the state variable x_i located at a point in the 3-dimensional space with spatial coordinate (α,β,γ). Observe that the Reaction Diffusion CNN equation (4) consists of a system of $N = nN_\alpha N_\beta N_\gamma$ *ordinary differential equations* (ODE's).

We will henceforth refer to the process of transforming a *PDE* into a *CNN equation* as *mapping a PDE into a CNN*. Table 1 shows the mapping of 4 well-known reaction diffusion PDE's.

Table 1. Mapping Reaction-Diffusion PDE into a Reaction Diffusion CNN.

FitzHugh-Nagumo CNN Equation	*FitzHugh-Nagumo PDE*
$\dot{u}_i = -\left(\dfrac{u_i^3}{3} - u_i\right) - v_i$ $+D_1[u_{i+1} + u_{i-1} - 2u_i]$ $\dot{v}_i = -\epsilon[u_i - bv_i + a]$	$\dfrac{\partial u}{\partial t} = -\left(\dfrac{u^3}{3} - u\right) - v + D_1\dfrac{\partial^2 u}{\partial x^2}$ $\dfrac{\partial v}{\partial t} = -\epsilon[u - bv + a]$
Brusselator CNN Equation	*Brusselator PDE*
$\dot{u}_{ij} = a - (b+1)u_{ij} + u_{ij}^2 v_{ij} + D_1[u_{i+1,j}$ $+u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij}]$ $\dot{v}_{ij} = bu_{ij} - u_{ij}^2 v_{ij} + D_2[v_{i+1,j}$ $+v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{ij}]$	$\dfrac{\partial u}{\partial t} = a - (b+1)u + u^2 v + D_1\left[\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right]$ $\dfrac{\partial v}{\partial t} = bu - u^2 v + D_2\left[\dfrac{\partial^2 v}{\partial x^2} + \dfrac{\partial^2 v}{\partial y^2}\right]$
Meinhardt-Gierer CNN Equation	*Meinhardt-Gierer PDE*
$\dot{u}_{ij} = \dfrac{\alpha u_{ij}^2}{v_{ij}} - \beta u_{ij} + D_1[u_{i+1,j} + u_{i-1,j}$ $+u_{i,j+1} + u_{i,j-1} - 4u_{ij}]$ $\dot{v}_{ij} = \alpha u_{ij}^2 - \gamma v_{ij} + D_2[v_{i+1,j} + v_{i-1,j}$ $+v_{i,j+1} + v_{i,j-1} - 4v_{ij}]$	$\dfrac{\partial u}{\partial t} = \dfrac{\alpha u^2}{v} - \beta u + D_1\left[\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right]$ $\dfrac{\partial v}{\partial t} = \alpha u^2 - \gamma v + D_2\left[\dfrac{\partial^2 v}{\partial x^2} + \dfrac{\partial^2 v}{\partial y^2}\right]$
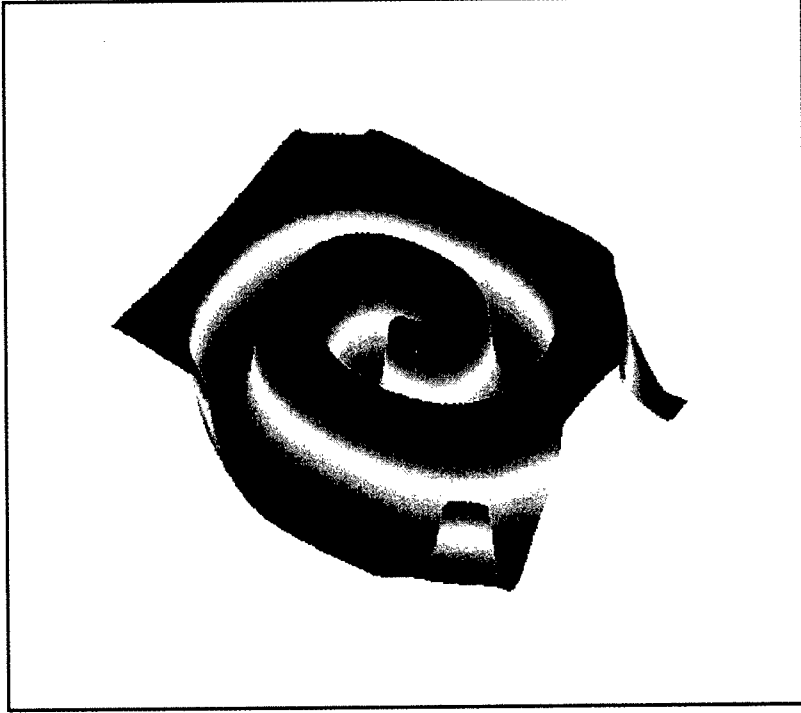
Figure 6. Three-dimensional view of a steadily rotating spiral wave at one instant of time. The vertical axis (perpendicular to the plane) represents the capacitor voltage across Chua's diode and the horizontal axes (on the plane) are the spatial coordinates.

Oregonator CNN Equation	Oregonator PDE
$\epsilon \dot{u}_{ijk} = u_{ijk} + v_{ijk} - \alpha u_{ijk}^2 - u_{ijk}v_{ijk} +$ $D_1[u_{i+1,j,k} + u_{i-1,j,k} + u_{i,j+1,k}$ $+u_{i,j-1,k} + u_{i,j,k+1} + u_{i,j,k-1}$ $-6u_{ijk}]$	$\epsilon \dfrac{\partial u}{\partial t} = u + v - \alpha u^2 - uv$ $+D_1 \left[\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2} \right]$
$\dot{v}_{ijk} = -v_{ijk} + \beta w_{ijk} - u_{ijk}v_{ijk} +$ $D_2[v_{i+1,j,k} + v_{i-1,j,k} + v_{i,j+1,k}$ $+v_{i,j-1,k} + v_{i,j,k+1} + v_{i,j,k-1}$ $-6v_{ijk}]$	$\dfrac{\partial v}{\partial t} = -v + \beta w - uv$ $+D_2 \left[\dfrac{\partial^2 v}{\partial x^2} + \dfrac{\partial^2 v}{\partial y^2} + \dfrac{\partial^2 v}{\partial z^2} \right]$
$\dot{w}_{ijk} = u_{ijk} - w_{ijk} + D_3[w_{i+1,j,k}$ $+w_{i-1,j,k} + w_{i,j+1,k} + w_{i,j-1,k}$ $+w_{i,j,k+1} + w_{i,j,k-1} - 6w_{ijk}]$	$\dfrac{\partial w}{\partial t} = u - w$ $+D_3 \left[\dfrac{\partial^2 w}{\partial x^2} + \dfrac{\partial^2 w}{\partial y^2} + \dfrac{\partial^2 w}{\partial z^2} \right]$

Reaction diffusion PDE's form an important, albeit relatively small, subset of the universe of all nonlinear PDE's. For any other nonlinear PDE, the *spatial derivative* (of any order) of any *state variable* $u_i(t; x, y, z)$ at *time t* and *spatial location* [d] $(x, y, z) \in \mathbf{R}^3$ can be *approximated* to any desired accuracy [24,25] by *finite differences* involving only the values of the state variable $u_i(t; \alpha', \beta', \gamma')$ located at a *finite number* of lattice points $(x, y, z) = (\alpha', \beta', \gamma') \in S_{\alpha\beta\gamma}$, where $(\alpha, \beta, \gamma) \in \mathbf{Z}^3$ denotes the "discrete" coordinates of a 3-dimensional lattice $\mathbf{Z}^3$, and $S_{\alpha\beta\gamma}(r)$ denotes a neighborhood of radius r, centered at (α, β, γ). Hence, given any nonlinear PDE, we can generate many (depending on the desired accuracy) approximate discretized systems of ODE's in terms of a 3-dimensional array of state variables $x_i(\alpha, \beta, \gamma)$, $\alpha = 1, 2, ..., N_\alpha$, $\beta = 1, 2, ..., N_\beta$, $\gamma = 1, 2, ..., N_\gamma$. Moreover, since all *finite-difference* operations involve only variables located within a *local* neighborhood, we can always *decompose*[e] the discretized system into a component (the *isolated cell*) which involves only $u_i(t; \alpha, \beta, \gamma)$ at the lattice site (α, β, γ), and another component(the *cell coupling*) which involves all neighboring cells $(\alpha', \beta', \gamma') \in S_{\alpha\beta\gamma}(r)$. In other words, given any nonlinear PDE, we can induce many *associated CNN equations*— the examples given in *Table 1* represent the simplest examples. Although it is *not true* that the qualitative behaviors of a nonlinear PDE and its associated CNN equation are always the same—the propagation failure phenomenon[26] is a case in point, extensive computer experiments have shown that for the vast majority of cases, the respective solutions can be made virtually indistinguishable by choosing a sufficiently large array size and by optimizing the CNN cell and coupling parameters [22].

It is important to observe that partial differential equations are merely mathematical abstractions of nature. The concept of a *continuum* is in fact an idealization of reality. Even the collection of all electrons in a solid does not form a continuum because much of the volume separating the electrons from the nucleus of atoms[f] represents a vast empty space! In fact, recent works by Smolin and his colleagues have proved that *"the spectrum of the volume of any physical region is discrete"*[27]. In particular, quantum mechanics implies that at extremely small distances even space is made of discrete bits!

3 Local Activity Is the Origin of Complexity

Let $\mathcal{N}$ be a 2-dimensional CNN associated with a homogeneous medium. By definition, a CNN is defined uniquely by *cells* and their *interactions*. Let us identify each cell $C(j, k)$ as a nonlinear *m-port* defined by its cell dynamics, where m is equal to the number of state variables which are *directly coupled* to its neighbors, as depicted by the "m" external (ungrounded) terminals attached to each cell $C(j, k)$ in Fig.7 for a reaction diffusion CNN. Note that the cell $C(j, k)$ may contain additional state

---

[d]To avoid clutter we restrict our discussion to the 3-dimensional Euclidean space $\mathbf{R}^3$. The same formulation is valid for any dimension.

[e]In most cases, the decomposition consists of just the superposition of these two components. However, more complex decompositions (e.g., nonlinear functional compositions) may be required for some nonlinear PDE's.

[f]The ratio between the diameter of the electron orbits in an atom to the diameter of its nucleus is of the order of 10^5.

variables which are not directly coupled to its neighbors and hence are suppressed in Fig.7 to avoid clutter.[9]

Let us identify next the *couplings* from cell $C(j,k)$ to all cells within the 3×3 sphere of influence S_{jk} centered at (j,k) by a γ-port Γ_{jk}. For the one-diffusion reaction diffusion CNN shown in Fig.5, an isolated cell $C(j,k)$ and its 4 coupling resistors are shown in Fig.8(a). Note that the 5-terminal coupling circuit, or γ-port Γ_{jk} ($\gamma = 4$ in this case), in Fig.8(a) can be redrawn into the form of a *grounded* 4-port (i.e., with a common ground terminal) in Fig.8(b). For the reaction-diffusion CNN depicted in Fig.7, the γ-port Γ_{jk} is composed of m identical grounded 4-ports(with node (j,k) as the ground node) made of 4 identical positive resistances, so that $\gamma = 4m$. Observe that a CNN is completely specified by the *m-port cell* $C(j,k)$ and the γ-port coupling Γ_{jk} since they can be used as a template to build up a CNN of any array size.

Since the conductance of all resistors in each layer "i" of the resistive grid in Fig.7 is equal to the diffusion coefficient D_i, which is assumed to be positive, it follows that the γ-port coupling Γ_{jk} is passive. If the cell $C(j,k)$ is *not locally active*, then it follows from symmetry considerations and the qualitative theory of nonlinear networks[28,29,30,31] that the CNN must have a *unique* steady state solution, thereby implying that all nodes belonging to the same layer must have identical node-to-datum voltages. It follows that no patterns or dissipative structures can exist.

In the general case, the γ-port coupling Γ_{jk} may consist of a *nonlinear dynamical* multi-port. In this case, the dynamics must again tend to a *homogeneous* node voltage distribution on each resistor grid if both the m-port cell $C(j,k)$ *and* the Γ-port coupling Γ_{jk} are *not locally active*. The above analysis justifies the following fundamental result on complexity:

The Local Activity Principle

> A CNN associated with a homogeneous non-conservative (i.e., *not lossless*) medium having a *zero-flux* boundary condition can *not* exhibit patterns or dissipative structures unless the *cells*, or the *couplings*, are *locally active*.

4 Local Activity for Reaction-Diffusion CNNs

In general, each cell $C(j,k,l)$ in a reaction-diffusion CNN has n state variables but only $m \leq n$ among them are coupled directly to its nearest neighbors via diffusion.

---

[9]A state variable x_k is suppressed if its associated diffusion coefficient D_k is zero.

12

In this case, the state equation of each cell $C(j, k, l)$ assumes the form

$$
\left.
\begin{aligned}
\dot{V}_1(j, k, l) &= f_1(V_1(j, k, l), V_2(j, k, l), ..., V_n(j, k, l)) + D_1 \nabla^2 V_1(j, k, l) \\
\dot{V}_2(j, k, l) &= f_2(V_1(j, k, l), V_2(j, k, l), ..., V_n(j, k, l)) + D_2 \nabla^2 V_2(j, k, l) \\
&\vdots \\
\dot{V}_m(j, k, l) &= f_m(V_1(j, k, l), V_2(j, k, l), ..., V_n(j, k, l)) + D_m \nabla^2 V_m(j, k, l) \\
\dot{V}_{m+1}(j, k, l) &= f_{m+1}(V_1(j, k, l), V_2(j, k, l), ..., V_n(j, k, l)) \\
&\vdots \\
\dot{V}_n(j, k, l) &= f_n(V_1(j, k, l), V_2(j, k, l), ..., V_n(j, k, l)) \\
j &= 1, 2, ..., N_x; k = 1, 2, ..., N_y; l = 1, 2, ..., N_z.
\end{aligned}
\right\}
\quad (5)
$$

Alternatively, we can include the remaining Laplacian terms $D_{m+1} \nabla^2 V_{m+1}(j, k, l)$, $D_{m+2} \nabla^2 V_{m+2}(j, k, l)$, ..., $D_n \nabla^2 V_n(j, k, l)$ in Eq.(5) and set $D_{m+1} = D_{m+2} = ... = D_n = 0$. Recasting Eq.(5) into vector form, we obtain

$$\dot{\mathbf{V}}_a = \mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b) + \mathbf{D}_a \nabla^2 \mathbf{V}_a \qquad (6)$$

$$\dot{\mathbf{V}}_b = \mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b) \qquad (7)$$

where

$$\mathbf{V}_a = [V_1(j, k, l), V_2(j, k, l), ...V_m(j, k, l)]^T \qquad (8)$$

$$\mathbf{V}_b = [V_{m+1}(j, k, l), V_{m+2}(j, k, l), ...V_n(j, k, l)]^T \qquad (9)$$

$$\mathbf{f}_a = [f_1(\cdot), f_2(\cdot), ...f_m(\cdot)] \qquad (10)$$

$$\mathbf{f}_b = [f_{m+1}(\cdot), f_{m+2}(\cdot), ...f_n(\cdot)] \qquad (11)$$

$$\nabla^2 \mathbf{V}_a = [\nabla^2 V_1(j, k, l), \nabla^2 V_2(j, k, l), ..., \nabla^2 V_m(j, k, l)] \qquad (12)$$

$$
\mathbf{D}_a \triangleq
\begin{bmatrix}
D_1 & & & \\
& D_2 & & \\
& & \ddots & \\
& & & D_m
\end{bmatrix}, D_i > 0
\qquad (13)
$$

To emphasize that a CNN is defined by specifying the dynamics of the *cells* and their *couplings (cell interaction laws)*, let us rewrite Eqs.(6)-(7) into the following standard form [32]:

cell dynamics:

$$\dot{\mathbf{V}}_a = \mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b) + \mathbf{I}_a \qquad (14)$$

$$\dot{\mathbf{V}}_b = \mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b) \qquad (15)$$

cell interaction laws:

$$
\mathbf{I}_a = \mathbf{g}_a\Big(\mathbf{V}_a(j,k,l), \mathbf{V}_a(j-1,k,l), \mathbf{V}_a(j+1,k,l), \mathbf{V}_a(j,k-1,l), \mathbf{V}_a(j,k+1,l),
$$

$$
\mathbf{V}_a(j,k,l-1), \mathbf{V}_a(j,k,l+1)\Big)
$$

$$
= \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_m \end{bmatrix} \begin{bmatrix} \nabla^2 V_1(j,k,l) \\ \nabla^2 V_2(j,k,l) \\ \vdots \\ \nabla^2 V_m(j,k,l) \end{bmatrix} \tag{16}
$$

It is important to observe that Eqs.(14)-(15) defining the *cell dynamics* involve only the voltage variables $(\mathbf{V}_a(j,k,l), \mathbf{V}_b(j,k,l))$ and the current variables $\mathbf{I}_a(j,k,l)$ at *the same* spatial location (j,k,l),[h] whereas Eq.(16) defining the *cell interaction laws* involves *not only* the voltage variable $\mathbf{V}_a(j,k,l)$ at spatial location (j,k,l), but also those of the neighboring cells. The relationship between a typical *cell $C(j,k,l)$* at location (j,k,l) and its *coupling* network is shown in Fig.7 for a 2-dimensional CNN reaction diffusion equation, where $\mathbf{V}_a = [V_1(j,k), V_2(j,k), ..., V_m(j,k)]^T$ and $\mathbf{I}_a = [I_1(j,k), I_2(j,k), ..., I_m(j,k)]^T$ are the "m" port voltage and port current variables. In the special case when $m = 1$, Fig.7 reduces to Fig.8(a), where cell $C(j,k)$ degenerates into a one-port.

4.1   Cell Equilibrium Points

Let us define the *static characteristic* of an isolated CNN cell by setting $\dot{\mathbf{V}}_a = 0$ and $\dot{\mathbf{V}}_b = 0$ in Eqs.(14) and (15); namely,

$$
\mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b) + \mathbf{I}_a = 0 \tag{17}
$$

$$
\mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b) = 0 \tag{18}
$$

where $\mathbf{V}_a \in \mathbf{R}^m$, $\mathbf{I}_a \in \mathbf{R}^m$, $\mathbf{V}_b \in \mathbf{R}^{n-m}$, $\mathbf{f}_a \in \mathbf{R}^m$, $\mathbf{f}_b \in \mathbf{R}^{n-m}$. Solving Eq.(18) for $\mathbf{V}_b$ in terms of $\mathbf{V}_a$, we obtain

$$
\mathbf{V}_b = \mathbf{g}_a(\mathbf{V}_a) \tag{19}
$$

where $\mathbf{g}_a(\cdot)$ may be a *multi-valued* function of $\mathbf{V}_a$.

Substituting Eq.(19) for $\mathbf{V}_b$ in Eq.17, we obtain the following *implicit* and possibly *multi-valued* function:

Static Cell Characteristic
$$
\mathcal{G}(\mathbf{V}_a, \mathbf{I}_a) \triangleq \mathbf{f}_a(\mathbf{V}_a, \mathbf{g}_a(\mathbf{V}_a)) + \mathbf{I}_a = 0 \tag{20}
$$

---

[h]This observation allows us to suppress the spatial coordinates in Eqs.(14)-(15) without ambiguity.

Let us solve Eq.(20) for $\mathbf{V}_a$. In general, there can be many solutions $\mathbf{V}_a = \mathbf{V}_a(Q_1)$, $\mathbf{V}_a(Q_2)$, ..., $\mathbf{V}_a(Q_p)$ for *each* $\mathbf{I}_a = \bar{\mathbf{I}}_a \in \mathbf{R}^m$, where $\mathbf{V}_a(Q_i)$ denotes the port voltage solution $\mathbf{V}_a$ at the *ith* solution Q_i. We call each solution $\mathbf{V}_a = \mathbf{V}_a(Q_i)$ a *cell equilibrium point* associated with $\mathbf{I}_a = \bar{\mathbf{I}}_a \in \mathbf{R}^m$. In other words, each cell equilibrium point $\mathbf{V}_a(Q_i)$ is *parameterized* by $\mathbf{I}_a \in \mathbf{R}^m$. The loci of all such cell equilibrium points calculated *explicitly* at $\mathbf{I}_a = \bar{\mathbf{I}}_a$, as $\bar{\mathbf{I}}_a$ ranges over the entire m-dimensional Euclidean space $\mathbf{R}^m$ is identical to the static *cell characteristic* defined *implicitly* in Eq.(20).

4.2 Cell Complexity Matrix

Let Q_i be a cell equilibrium point associated with $\mathbf{I}_a = \mathbf{I}_a(Q_i)$. Let $\mathbf{V}_a(Q_i)$ (obtained from Eqs.(20)) be the corresponding cell equilibrium point of the cell state equations Eqs.(14)-(15). Let

$$\mathbf{J}(Q_i) \triangleq \begin{bmatrix} \mathbf{A}_{aa}(Q_i) & \mathbf{A}_{ab}(Q_i) \\ \mathbf{A}_{ba}(Q_i) & \mathbf{A}_{bb}(Q_i) \end{bmatrix} \tag{21}$$

denote the $m \times m$ Jacobian matrix associated with $\mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b)$ and $\mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b)$, evaluated at $(\mathbf{V}_a(Q_i), \mathbf{V}_b(Q_i))$, where $\mathbf{V}_b(Q_i) = \mathbf{g}_a(\mathbf{V}_a)$ from Eq.(19), namely,

Small Signal CNN Cell Coefficients at Q_i

$$\mathbf{A}_{aa}(Q_i) = \frac{\partial \mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b)}{\partial \mathbf{V}_a}\bigg|_{\mathbf{V}_a=\mathbf{V}_a(Q_i), \mathbf{V}_b=\mathbf{V}_b(Q_i)},$$

$$\mathbf{A}_{ab}(Q_i) = \frac{\partial \mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b)}{\partial \mathbf{V}_b}\bigg|_{\mathbf{V}_a=\mathbf{V}_a(Q_i), \mathbf{V}_b=\mathbf{V}_b(Q_i)},$$

$$\mathbf{A}_{ba}(Q_i) = \frac{\partial \mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b)}{\partial \mathbf{V}_a}\bigg|_{\mathbf{V}_a=\mathbf{V}_a(Q_i), \mathbf{V}_b=\mathbf{V}_b(Q_i)},$$

$$\mathbf{A}_{bb}(Q_i) = \frac{\partial \mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b)}{\partial \mathbf{V}_b}\bigg|_{\mathbf{V}_a=\mathbf{V}_a(Q_i), \mathbf{V}_b=\mathbf{V}_b(Q_i)}. \tag{22}$$

Let $N(Q_i)$ be the linearized CNN cell at Q_i associated with the state Equations (14)-(15) of the m-port cell $C(j, k, l)$ obtained by deleting the higher order terms in the Taylor series expansion of $\mathbf{f}_a(\mathbf{V}_a, \mathbf{V}_b)$ and $\mathbf{f}_b(\mathbf{V}_a, \mathbf{V}_b)$ about Q_i; namely,

Linearized CNN Cell dynamics at Q_i

$$\dot{\mathbf{v}}_a = \mathbf{A}_{aa}\mathbf{v}_a + \mathbf{A}_{ab}\mathbf{v}_b + \mathbf{i}_a \tag{23}$$

$$\dot{\mathbf{v}}_b = \mathbf{A}_{ba}\mathbf{v}_a + \mathbf{A}_{bb}\mathbf{v}_b \tag{24}$$

where $\mathbf{v}_a \triangleq \mathbf{V}_a - \mathbf{V}_a(Q_i)$, $\mathbf{v}_b \triangleq \mathbf{V}_b - \mathbf{V}_b(Q_i)$, $\mathbf{i}_a \triangleq \mathbf{I}_a - \mathbf{I}_a(Q_i)$ are the infinitesimal voltages and currents, respectively, about the equilibrium point Q_i.

Definition 1: Local Activity

A CNN cell $C(j,k,l)$ is said to be *locally active* at Q_i iff there exists some $i_a(t)$ and some time $T > 0$ such that

$$\delta\mathcal{E}(Q_i) \triangleq \int_0^T < \mathbf{v}_a(t),\ i_a(t) > dt < 0 \qquad (25)$$

where $< \cdot, \cdot >$ denotes the vector dot product, and $\mathbf{v}_a(t)$ is the solution obtained by solving Eqs.(23)-(24) under *zero initial state* $\mathbf{v}_a(0) = \mathbf{0}$ and $\mathbf{v}_b(0) = \mathbf{0}$.

To derive a test for local activity, let us take the Laplace transform of Eq.(24) to obtain

$$s\hat{\mathbf{v}}_{\mathbf{a}}(s) = \mathbf{A}_{aa}\hat{\mathbf{v}}_{\mathbf{a}}(s) + \mathbf{A}_{ab}\hat{\mathbf{v}}_{\mathbf{b}}(s) + \hat{\mathbf{i}}_{\mathbf{a}}(s) \qquad (26)$$

$$s\hat{\mathbf{v}}_{\mathbf{b}}(s) = \mathbf{A}_{ba}\hat{\mathbf{v}}_{\mathbf{a}}(s) + \mathbf{A}_{bb}\hat{\mathbf{v}}_{\mathbf{b}}(s) \qquad (27)$$

where $\hat{\mathbf{v}}_{\mathbf{a}}(s)$, $\hat{\mathbf{v}}_{\mathbf{b}}(s)$ and $\hat{\mathbf{i}}_{\mathbf{a}}(s)$ denote the Laplace transform of $\mathbf{v}_a(t)$, $\mathbf{v}_b(t)$, and $i_a(t)$, respectively. Solving for $\hat{\mathbf{v}}_{\mathbf{b}}(s)$ from Eq.(27), we obtain

$$\hat{\mathbf{v}}_b(s) = (s\mathbf{1} - \mathbf{A}_{bb})^{-1}\mathbf{A}_{ba}\hat{\mathbf{v}}_a(s) \qquad (28)$$

Substituting Eq.(28) for $\hat{\mathbf{v}}_b(s)$ in Eq.(26) and solving for $\hat{\mathbf{i}}_a(s)$, we obtain

$$\hat{\mathbf{i}}_a(s) = \mathbf{Y}_Q(s)\mathbf{v}_a(s) \qquad (29)$$

where

CNN Cell Complexity Matrix at Q_i

$$\mathbf{Y}_Q(s) \triangleq (s\mathbf{1} - \mathbf{A}_{aa}) - A_{ab}(s\mathbf{1} - \mathbf{A}_{bb})^{-1}\mathbf{A}_{ba} \qquad (30)$$

is called the *CNN cell complexity matrix* at the cell equilibrium point Q_i. It follows from a classic theorem in *circuit theory* [33] that for a *reaction diffusion CNN Cell* to be *locally active* at Q_i, $\mathbf{Y}_Q(s)$ should *not* be a *positive real matrix* at Q_i. Hence, in order for a reaction diffusion CNN equation to exhibit complexity, it is *necessary* that the cell parameters be chosen such that the *cell complexity matrix* $\mathbf{Y}_Q(s)$ is *not* positive real at Q_i. The mathematical conditions for testing this local activity property is as follows:

Local activity criteria for reaction diffusion CNN

A Reaction Diffusion CNN Cell is *locally active* at a cell equilibrium point Q_i if, and only if, its *cell complexity matrix* $\mathbf{Y}_Q(s)$, or its *inverse cell complexity matrix* $\mathbf{Z}_Q(s) \triangleq \mathbf{Y}_Q^{-1}(s)$ (In this case, simply change the symbol from $\mathbf{Y}$ to $\mathbf{Z}$ in the following 4 conditions) satisfies *any one* of the following 4 conditions:

1. $\mathbf{Y}_Q(s)$ has a *pole* in $Re[s] > 0$.

2. $\mathbf{Y}_Q^H(i\omega) \triangleq \mathbf{Y}_Q^{\dagger}(i\omega) + \mathbf{Y}_Q(i\omega)$ is *not* a *positive semi-definite* matrix at some $\omega = \omega_0$, where ω_0 is any real number, and $\dagger$ denotes the Hermitian operator.

3. $\mathbf{Y}_Q(s)$ has a *simple* pole $s = i\omega_p$ on the *imaginary axis* where its associated *residue* matrix

$$\mathbf{K}_{-1} \triangleq \left. \begin{array}{l} \lim_{s \to i\omega_p}(s - i\omega_p)\mathbf{Y}_Q(s), \; if \; \omega_p < \infty \\ \lim_{\omega_p \to \infty} \frac{\mathbf{Y}(i\omega_p)}{i\omega_p}, \qquad if \; \omega_p = \infty \end{array} \right\}$$

is either *not* a Hermitian matrix, or else *not a positive semi-definite Hermitian* matrix.

4. $\mathbf{Y}_Q(s)$ has a *multiple* pole on the imaginary axis.

4.3 Reaction Diffusion CNN: One Diffusion Coefficient

Many well-known reaction diffusion CNNs have only one non-zero diffusion coefficient and 2 state variables, i.e., $m = 1$, $n = 2$. The most famous example belonging to this class is the FitzHugh Nagumo CNN Equation given in *Table 1*. For this class of reaction diffusion CNN, all variables in Eq.(24) are scalars and let us rewrite it using the established notation [22] for ease of reference:

$$\dot{v}_1 = a_{11}v_1 + a_{12}v_2 + i_1$$
$$\dot{v}_2 = a_{21}v_1 + a_{22}v_2 \tag{31}$$

The inverse *CNN cell complexity matrix* $Y_Q^{-1}(s) \triangleq Z_Q(s) = \frac{\dot{v}_1(s)}{i_1(s)}$ associated with Eq.(31) is a 1×1 matrix, or scalar function of s in this case, and is given by

$$Z_Q(s) = \frac{(s - a_{22})}{s^2 - Ts + \Delta} \tag{32}$$

where

$$T = a_{11} + a_{22} \tag{33}$$

$$\Delta = a_{11}a_{22} - a_{12}a_{21} \tag{34}$$

are the *trace* and *determinant* of the associated Jacobian matrix

$$\mathbf{J}(Q_i) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{35}$$

evaluated at the cell equilibrium point Q_i.

Applying the above *local activity criteria* to this special scalar case (with $Y_Q(s)$ replaced by $Z_Q(s)$), we obtain the following useful corollary:

Local Activity Criteria for One Diffusion Coefficient A one-port Reaction-Diffusion CNN cell with one diffusion coefficient and two state variables is *locally active* at a cell equilibrium point $Q = (\bar{V}_1, \bar{I}_1)$ if, and only if, any one of the following 4 conditions holds at Q:

1. $a_{11} + a_{22} > 0$ or $a_{11}a_{22} < a_{12}a_{21}$.

2. $a_{11} > 0$, or $a_{11} > \frac{a_{12}a_{21}}{a_{22}}$, if $a_{11} \leq 0$ and $a_{22} \neq 0$.

3. $a_{11}a_{22} > a_{12}a_{21}$, $a_{11} + a_{22} = 0$ and $a_{22} \neq 0$.

4. $a_{11}a_{22} = a_{12}a_{21}$, $a_{11} + a_{22} = 0$, and $a_{22} \neq 0$.

$$(36)$$

Although the above local activity criteria is couched in terms of the 4 Jacobian coefficients a_{11}, a_{12}, a_{21} and a_{22}, the criteria can be recast into the following 4 equivalent conditions involving only the 3 parameters Δ, T, and a_{22}:

Equivalent condition 1: $T > 0$, *or*
$$\Delta < 0$$

Equivalent condition 2: $T > a_{22}$, *or*
$$T \leq a_{22} \text{ and } a_{22}\Delta > 0$$

Equivalent condition 3: $T = 0$, *and*
$$\Delta > 0 \text{ and}$$
$$a_{22} \neq 0$$

$$(37)$$

Equivalent condition 4: $T = 0$, *and*
$$\Delta = 0 \text{ and}$$
$$a_{22} = 0$$

To visualize the regions in the Δ-T-a_{22} Euclidean space represented by the above system of inequalities, it is more convenient to consider a Δ-T-a_{22} cylindrical subspace and partition it into 8 uniform wedges above $\Delta = 0$, and 8 uniform wedges below $\Delta = 0$, as depicted in Fig.9. In terms of the Δ-T-a_{22} cylinder, *Equivalent condition 1* is represented by all points (a_{22}, T, Δ) *behind* the vertical separating plane $T = 0$, and all points below the horizontal separating plane $\Delta = 0$. *Equivalent condition 2* is represented by the wedges labeled $\boxed{1}$, $\boxed{2}$, $\boxed{3}$, $\boxed{4}$, $\boxed{5}$, $\boxed{7}$, and $\boxed{8}$ in the *upper half* ($\Delta > 0$)
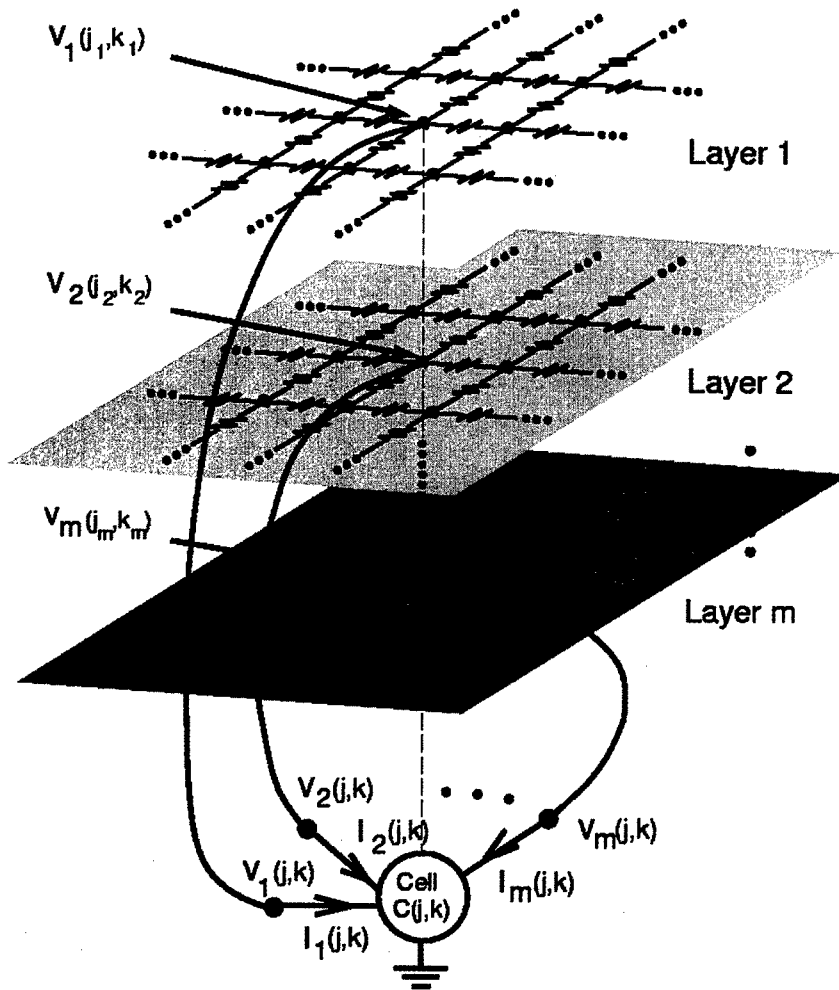
Figure 7. A 2-dimensional reaction-diffusion CNN having "m" non-zero diffusion coefficients ($D_i \neq 0, i = 1, 2, ..., m$). The subscript "$i$" attached to ($j_i, k_i$) denotes the *node* in the *ith* resistive grid layer which is connected to a terminal of cell $C(j, k)$. All resistors in layer i are linear resistors with identical conductance equal to D_i Siemens, $i = 1, 2, ..., m$.
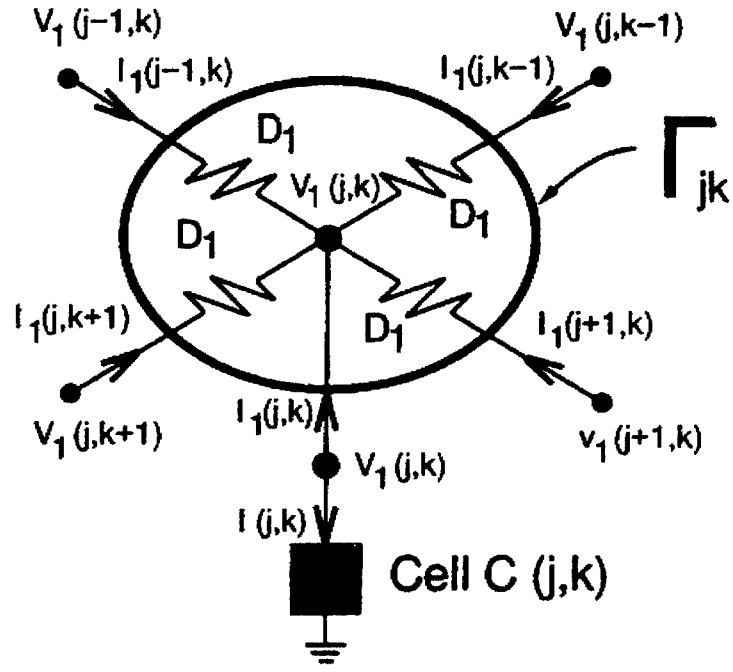
(a)
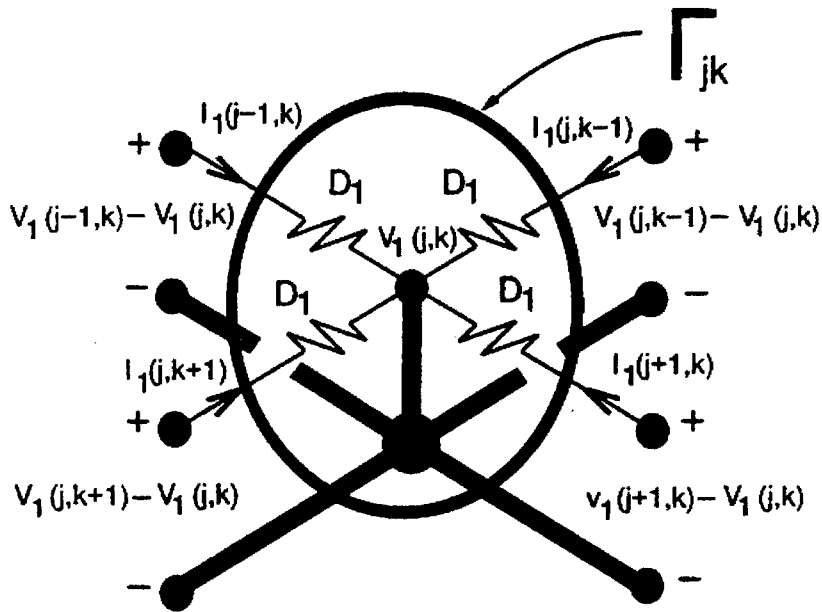
(b)

Figure 8. (a) An isolated cell $C(j,k)$ from Fig.5 with its 4 coupling conductances D_1. (b) The 5-terminal circuit Γ_{jk} in (a) is equivalent to a 4-port where the bold wires form the common "ground" terminal of each port.

cylinder. *Equivalent condition 3* is represented by the vertical separating plane(excluding the vertical axis) in the upper half ($\Delta > 0$) cylinder. *Equivalent condition 4* is represented by a single point at the origin. It follows from this partitioned cylinder that a reaction diffusion CNN with one diffusion coefficient and two state variables is *locally active* at a cell equilibrium point Q if, and only if, its associated parameter $(a_{22}(Q), T(Q), \Delta(Q))$ at Q lies *outside* of the "blue" *sector* $\boxed{6}$ in the upper half cylinder.

4.4 Reaction Diffusion CNN: Two Diffusion Coefficients

Consider next the class of reaction diffusion CNNs with two diffusion coefficients ($D_1 > 0$ and $D_2 > 0$) and two state variables, i.e., $m = n = 2$. Both the Brusselator and the Meinhardt-Gierer CNN equations in *Table 1* belong to this class. In this case, the CNN cell is a 3-terminal (including the ground reference terminal), or *2-port*, device whose linearized cell dynamics about an equilibrium point Q_i is described by

$$\dot{v}_1 = a_{11}v_1 + a_{12}v_2 + i_1$$
$$\dot{v}_2 = a_{21}v_1 + a_{22}v_2 + i_2 \tag{38}$$

where a_{11}, a_{12}, a_{21}, and a_{22} are the small-signal cell coefficients at Q_i defined in Eq.(35). Applying the *local activity criteria* from section 4.2 to the associated 2×2 cell complexity matrix $\mathbf{Y}_Q(s)$ at Q_i, we obtain the following corollary:

Local activity criteria for two diffusion coefficients

A two-port reaction-diffusion CNN cell with two diffusion coefficients and two state variables is *locally active* at a cell equilibrium point $Q = (\bar{V}_1, \bar{V}_2, \bar{I}_1, \bar{I}_2)$ if, and only if, any one of the following two conditions holds at Q:

$$\boxed{\begin{array}{l} 1.\ a_{22} > 0 \\[2em] 2.\ 4a_{11}a_{22} < (a_{12} + a_{21})^2 \end{array}} \tag{39}$$

For a detailed application of the various local activity criteria presented in Sections 4.3 and 4.4 to concrete examples; namely, the FitzHugh-Nagumo Equation, the Brusselator Equation, and the Gierer-Mainhardt Equation, the reader is referred to [34], [35], and [36], respectively.

5 Concluding Remarks

The preceding analysis can be generalized to any homogeneous media which can be mapped to a CNN defined by *any* cell dynamics, and *any* cell interaction laws, not necessarily of the reaction diffusion type presented in Section 4. In particular, the coupling γ-port Γ_{jkl} can be any *nonlinear dynamical* multi-port. In such cases, in order for the CNN to exhibit *complexity*, either the m-port cell C_{ijk}, or the

$$\Delta = a_{11}a_{22} - a_{12}a_{21}$$

Vertical separating plane $T = 0$

$\Delta > 0$ $a_{22} < T < 0$

$\Delta > 0$ $T < 0, a_{22} > 0$

$\Delta > 0$ $a_{22} < T < 0$

$T = a_{11} + a_{22}$

Horizontal separating plane $\Delta = 0$

a_{22}

$a_{22} = T = \Delta = 0$
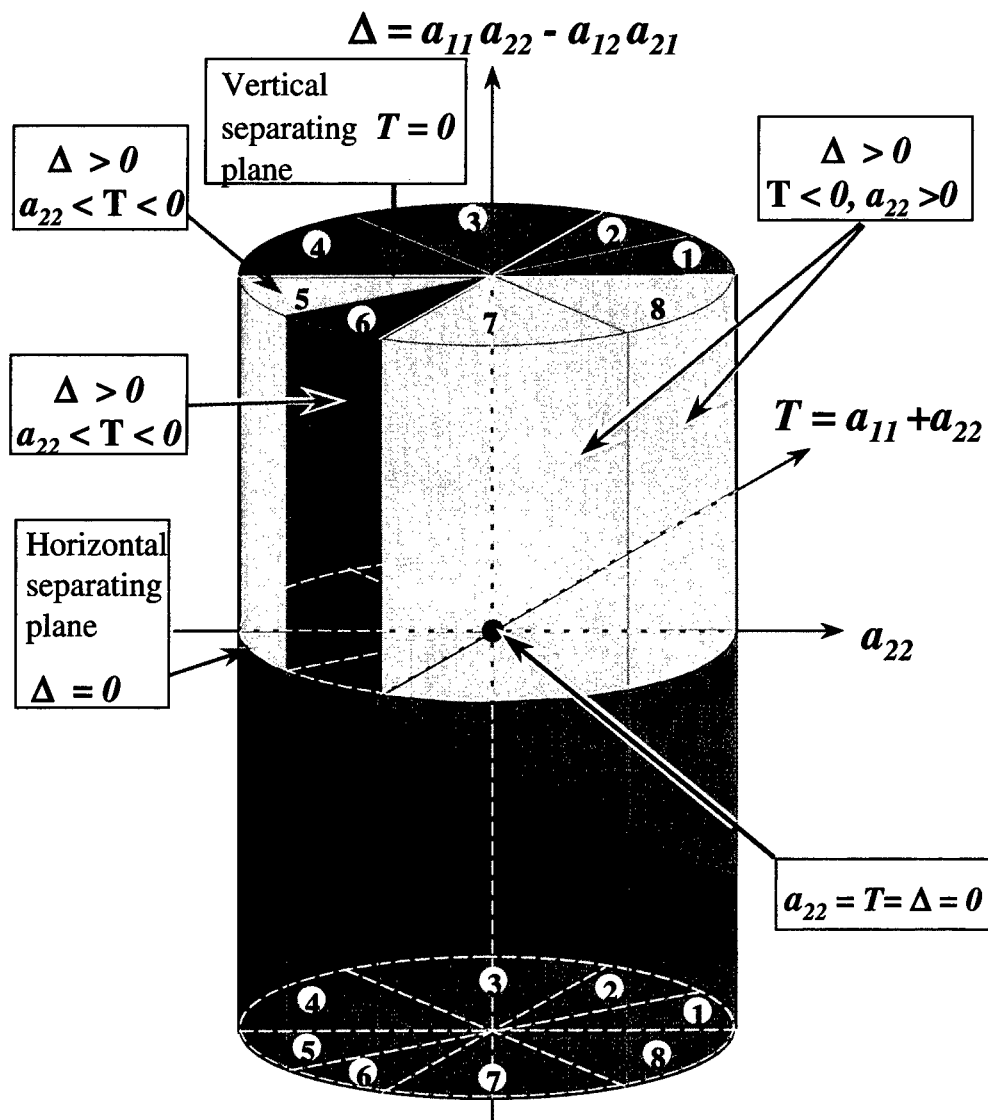
Figure 9. The Δ-T-a_{22} cylinder depicting the *locally passive* region represented by wedge $\boxed{6}$ in the upper half cylinder. All other regions are *locally active*. In particular, regions $\boxed{5}$, $\boxed{7}$, and $\boxed{8}$ in the upper half cylinder $\Delta > 0$ correspond to the *edge of chaos* where most complexities emerged.

coupling γ-port Γ_{ijk}, must be *locally active* at some *equilibrium point* of the *isolated* (disconnected from the rest of the CNN) *cell C_{ijk}*, or isolated *γ-port coupling Γ_{jkl}*.

While the CNN paradigm is an example of REDUCTIONISM *par excellence*, the true origin of emergence and complexity is traced to a much deeper new concept called *local activity*. The numerous complex phenomena unified under this mathematically precise principle include *self organization, dissipative structures, synergetics, order from disorder, far-from-thermodynamic equilibrium, collective behaviors, edge of chaos*, etc.

The central theme of the *local activity dogma* [22] asserts that the somewhat fuzzy notions of "emergence" and "complexity", as well as their various metamorphosis, such as those cited above, can all be rigorously explained by a precise scientific paradigm abstracted mathematically from the principle of conservation of energy; namely, a *CNN* operating near the *edge of chaos* [34,37], where the cells are not only *locally active*, but also *linearly asymptotically stable*. In particular, constructive and explicit mathematical inequalities are given for identifying the region in the CNN parameter space where complex phenomena may emerge, as well as for localizing it further into a relatively small parameter domain called the edge of chaos (regions $\boxed{5}$, $\boxed{7}$, and $\boxed{8}$ in Fig.9) where the potential for emergence is maximized.

> *Consequently: he who wants to have right without wrong,*
> *Order without disorder,*
> *Does not understand the principles*
> *Of heaven and earth.*
> *He does not know how*
> *Things hang together.*
>
> *Chuang Tzu*

References

1. J. H. Holland, *Emergence : from chaos to order*, Reading, Mass.: Addison-Wesley, 1998.
2. I. Stewart, *Life's other secret : the new mathematics of the living world*, New York : John Wiley, 1998.
3. H. Haken, "Visions of Synergetics," Special Issue on *Visions of Nonlinear Science in the 21st Century, International Journal of Bifurcation and Chaos*, 7(1997), pp.1927-1951.
4. R. Badii and A. Politi, *Complexity : hierarchical structures and scaling in physics*, Cambridge ; New York : Cambridge University Press, 1997.
5. C. G. Langton(Ed.), *Artificial life : an overview*, Cambridge, Mass. : MIT Press, 1995.
6. P. Coveney and R. Highfield, *Frontiers of complexity : the search for order in a chaotic world*, New York : Fawcett Columbine, 1995.
7. M. Gell-Mann, *The quark and the jaguar : Adventures in the simple and the complex*, New York : W.H. Freeman, 1994.
8. F. Crick, *The astonishing hypothesis : the scientific search for the soul*, New

York : Scribner ; New York : Maxwell Macmillan International, 1994.

9. K. Mainzer, *Thinking in complexity : the complex dynamics of matter, mind, and mankind*, Berlin ; New York : Springer-Verlag, 1994.

10. M. Eigen, *Steps towards life : a perspective on evolution*, Oxford ; New York: Oxford University Press, 1992.

11. M. M. Waldrop, *Complexity: the emerging science at the edge of order and chaos*, New York : Simon & Schuster, 1992.

12. L. Nadel and D. L. Stein, *1990 lectures in complex systems : the proceedings of the 1990 Complex Systems Summer School*, v. 3. of series *Santa Fe Institute studies in the sciences of complexity. Lectures*, Redwood City, Calif. : Addison-Wesley, 1991.

13. S. Forrest(Ed.), *Emergent computation : self-organizing, collective, and cooperative phenomena in natural and artificial computing networks*, Cambridge, Mass. : MIT Press, 1991.

14. G. Nicolis and I. Prigogine, *Exploring complexity : an introduction*, New York : W.H. Freeman, 1989.

15. D. L. Stein(Ed.), *Lectures in the sciences of complexity : the proceedings of the 1988 Complex Systems Summer School held June-July 1988 in Santa Fe, New Mexico*, Redwood City, Calif. : Addison-Wesley Pub. Co., Advanced Book Program, 1989.

16. Paul Manneville, *Dissipative structures and weak turbulence*, Boston : Academic Press, 1990.

17. R. Kapral and K. Showalter (Eds.), *Chemical waves and patterns*, v. 10 of series *Understanding chemical reactivity*, Dordrecht ; Boston : Kluwer Academic Publishers, 1995.

18. D. Walgraef, *Spatio-temporal pattern formation : with examples from physics, chemistry, and materials science*, in series *Partially ordered systems*, New York : Springer, 1997.

19. L.O. Chua, "Guest Editorial," *Special Issue on Nonlinear waves, patterns and spatio-temporal chaos, IEEE Trans. on Circuits and Systems-I*, **42**(1995), pp.557-558, Oct. 1995.

20. L. O. Chua and Y. W. Sing, "A nonlinear lumped circuit model for Gunn Diodes," *Int. Journal of Circuit Theory and Applications*, **6**(1978), pp.375-408.

21. A. V. Holden, "Nonlinear Science—the impact of biology," Special Issue on *Visions of Nonlinear Science in the 21st Century, International Journal of Bifurcation and Chaos*, **7**(1997), pp.2075-2104.

22. L. O. Chua, *CNN: A Paradigm for Complexity*, Singapore : River Edge, N.J. : World Scientific, 1998.

23. R. N. Madan(Ed.), *Chua's circuit: A paradigm for chaos*, Singapore : River Edge, N.J. : World Scientific, 1993.

24. L.V. Kantorovich and V.I. Krylov, *Approximate methods of higher analysis*, New York, Interscience, 1964.

25. L.M. Milne-Thomson, *The calculus of finite differences*, London: Macmillan, 1960.

26. V. Perez-Munuzuri, V. Perez-Villar and L.O. Chua, "Propagation failure in lin-

ear arrays of Chua's circuits," *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, **2**(1992), pp.403-406.

27. C. Rovelli and L. Smolin, "Discreteness of area and volume in quantum gravity," *Nuclear Physics B*, **B442**(1995), pp.593-619.

28. L.O. Chua and D.N. Green, "Graph-theoretic properties of dynamic nonlinear networks," *IEEE Trans. on Circuits and Systems*, **CAS-23**(1976), pp.292-312.

29. L.O. Chua, "A qualitative analysis of the behavior of dynamic nonlinear networks: stability of autonomous networks," *IEEE Trans. on Circuits and Systems*, **CAS-23**(1976), pp.355-379.

30. L.O. Chua, "Dynamic nonlinear networks: state-of-the-art," *IEEE Trans. on Circuits and Systems*, **CAS-27**(1976), pp.1059-1087.

31. L.O. Chua, "Nonlinear circuits," *IEEE Trans. on Circuits and Systems*, **CAS-31**(1976), pp.69-87.

32. L.O. Chua, M. Hasler, G.S. Moschytz, J. Neirynck, "Autonomous cellular neural networks: a unified paradigm for pattern formation and active wave propagation," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, **42**(1995), pp.559-577.

33. Brian D. O. Anderson and S. Vongpanitlerd, *Network analysis and synthesis: a modern systems theory approach*, in *Prentice-Hall electrical engineering series. Prentice-Hall networks series*, Englewood Cliffs, N.J.,Prentice-Hall, 1973.

34. R. Dogaru and L.O. Chua, "Edge of chaos and local activity domain of FitzHugh-Nagumo equation," *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, **8**(1998), pp.211-257.

35. R. Dogaru and L.O. Chua, "Edge of chaos and local activity domain of the Brusselator CNN," *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, **8**(1998), pp.1107-1130.

36. R. Dogaru and L.O. Chua, "Edge of chaos and local activity domain of the Gierer-Meinhardt CNN," *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, **8**(1998), pp.2321-2340.

37. L. Min, K.R. Crounse, and L.O. Chua, "Analytical criteria for local activity and applications to the Oregonator CNN," *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, **10**(2000), in press.

On the Existence of Spatially Uniform Scaling Laws in the Climate System

A. A. TSONIS AND P. J. ROEBBER

Department of Mathematical Sciences, Atmospheric Sciences Group,
University of Wisconsin-Milwaukee, Milwaukee, WI 53201-0413
E-mail: aatsonis@csd.uwm.edu

J. B. ELSNER

Department of Geography, Florida State University, Tallahassee, FL 32306
E-mail: jbelsner@elsner.coss.fsu.edu

Scale invariance (scaling) in a time series of an observable quantity is a symmetry law which when it exists can provide unique insights about the process in question. It describes variability and transitions at all scales and is often a result of nonlinear dynamics. It is well known that the spectra of atmospheric and climatic variables possess considerable power at low frequencies. Since "red" spectra often associate with scaling processes, it is reasonable to suppose that a search for scaling laws in climatic data might be fruitful. Consequently, the search for scaling in such data over the past decade has produced some exciting ways to describe climate variability. In the past and lately, there has been a growing interest in the existence of uniform in space temporal scaling laws for observable properties of the climate system, since such a property would provide a common rule describing temporal variability everywhere on the globe. Here we show that in spatially extended systems, uniform in space scaling demands that global averages be time invariant. A corollary to this is that where global averages do exhibit temporal variability, as in our climate system, spatial variation in scaling properties is required.

1 Introduction

A scaling (fractal) process $y(t)$ satisfies the relationship $y(t) \overset{d}{=} \sigma^{-1} y(\lambda t)$ where $\overset{d}{=}$ indicates equality in distribution and $\sigma, \lambda > 0$. This relationship indicates that the statistical properties at time scale t are related to the statistical properties at time scale λt. Consequently, any moment of order k, μ'_k, satisfies the relation $\mu'_k(t) = \sigma^{-1} \mu'_k(\lambda t)$. It is easy to show that the power law $\mu'_k(t) = At^H$ with $H = \log\sigma/\log\lambda$ is a solution to the last equation (Triantafyllou *et al.*[1]).

Recently, new approaches based on the theory of random walks have been developed to elucidate scaling in time series (Tsonis and Elsner[2], Viswanathan *et al.*[3]). According to these approaches a time series $x(t)$ representing some observable (temperature, pressure, etc.) is mapped onto a random walk whose net displacement, $y(t)$, after t time steps is defined by the running sum $y(t) = \sum_{i=1}^{t} x(i)$. For any walk a suitable statistical quantity that characterizes the walk is the root mean square fluctuation about the average displacement $F^2(t) = \overline{[\Delta y(t)]^2} - \overline{[\Delta y(t)]}^2$, where $\Delta y(t) = y(t_0 + t) - y(t_0)$ and the bars indicate an average over all positions t_0 in the walk. The calculation of $F(t)$ can distinguish three types of behavior: 1) uncorrelated time series described by $F(t) \propto t^H$ with $H = 0.5$ as expected from the central limit theorem; 2) time series exhibiting positive long-range correlations described by $F(t) \propto t^H$ with $H > 0.5$; and 3) time series exhibiting

negative long-range correlations described by $F(t) \propto t^H$ with H<0.5. Markov processes with local correlations extending up to some scale also give H=0.5 for sufficiently large t. It is well known (Feder[4]) that the correlation function C(t) of future increments, y(t), with past increments, y(-t), is given by $C(t) = 2(2^{2H-1} - 1)$. For H=0.5, we have C(t)=0 as expected, but for H≠0.5 we have C(t)≠0 independent of t. This indicates infinitely long correlations and leads to a scale-invariance (scaling) associated with positive long-range correlations for H>0.5 (i.e. an increasing trend in the past implies an increasing trend in the future) and to scaling associated with negative long-range correlations for H<0.5 (i.e. an increasing trend in the past implies a decreasing trend in the future). Random walks with H≠0.5 are referred to as fractional Brownian motions (fBms). In theory the exponent H is related to the spectra of the y(t) function via a relation of the form $S(f) \propto f^{-(2H+1)}$ and to the spectra of the x(t) function via a relation of the form $S(f) \propto f^{-2H+1}$ where f is the frequency.

2. Analysis and results

Let us assume that records of some meteorological variable exist at a sufficiently large number of stations (m) evenly distributed over the globe. Let us further assume that all those stations obey the spatially uniform scaling law $F(t) \sim t^H$ with the same exponent H. If $x_j(t)$, j=1,m are the records of the stations, then the global (planetary) mean of those records, $x_g(t)$, is given by:

$$x_g(t) = \frac{1}{m} \sum_{j=1}^{m} x_j(t) \qquad (1)$$

It follows that the displacement of the random walk generated by the global mean record is:

$$y_g(t) = \sum_{i=1}^{t} x_g(i) = \sum_{i=1}^{t} \{ \frac{1}{m} \sum_{j=1}^{m} x_j(i) \} = \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{i=1}^{t} x_j(i) \right] \qquad (2)$$

The sum in the bold brackets is the displacement of the walk for a particular station. Since we have assumed that at each station the records obey the law $F(t) \sim t^H$ with the same H, then the outer sum is zero (as it represents the average displacement after n time steps of many random walkers with the same exponent). In this case the above equation reduces to $y_g(t)=0$ which will indicate that the global mean $x_g(t)$ is also zero at any time.

This theoretically derived result can be verified by simple computer simulations. Consider m stations at which some variable y has been observed and that this quantity scales with the law $F(t) \sim t^H$ with H=0.7 at all stations. For illustrative purposes, we have generated such a function y(t) for m stations by inverting power spectra of the form $f^{-(2H+1)}$. The formula used to generate y(t) functions for t=1, N is given by $y(t) = \sum_{k=1}^{N/2} [Ck^{-a}(\frac{2\pi}{N})^{1-a}]^{1/2} \cos(2\pi tk/N + \phi_k)$ where C is a constant, N is the sample size, ϕ_k are N/2 random phases uniformly distributed in $[0, 2\pi]$, and a=2H + 1 (Osborne and Provenzale[5]; Tsonis[6]). Then using $y(t) = \sum_{i=1}^{t} x(i)$ the time series x(t) for each station was produced. From all the available stations we then estimated the global mean $x_g(t)$ for two sample sizes m (Figure 1). For m=10 the global mean fluctuates

significantly above zero but as m increases the global mean tends at all times to zero. This result demonstrates the theoretical proof provided by equation (2) that a spatially uniform scaling law requires time invariance in the global mean. This behavior is manifestly counter to that of our climate system, which exhibits nonlinearities and variability at all time scales. Indeed, in a recent study it was shown that for the global temperature record the relation between $y_g(t)$ and t involves multiple temporal scaling regimes (Tsonis et al.[7]).

3. Discussion

To those familiar with the theory of random walks this result may not be surprising. Nevertheless, due to limitations in data and other shortcomings, applying these ideas to problems in physical sciences is often misguided and the wrong conclusions are drawn (Koscienly-Bunde et al.[8]). From the above, it follows that in spatially extended systems displaying variability at all time scales temporal scaling must vary in space. The spatial distribution of scaling must, in some way, reflect the dynamics of the system. For the climate system, spatial variation in scaling has been clearly demonstrated in a recent study of the 500 hPa height field, which is hydrostatically linked to the mean temperature of the lower troposphere (Tsonis et al.[9]). In this work, local scaling patterns were linked to specific properties of the atmospheric general circulation (baroclinic instability, storm tracks and persistence of circulation anomalies).
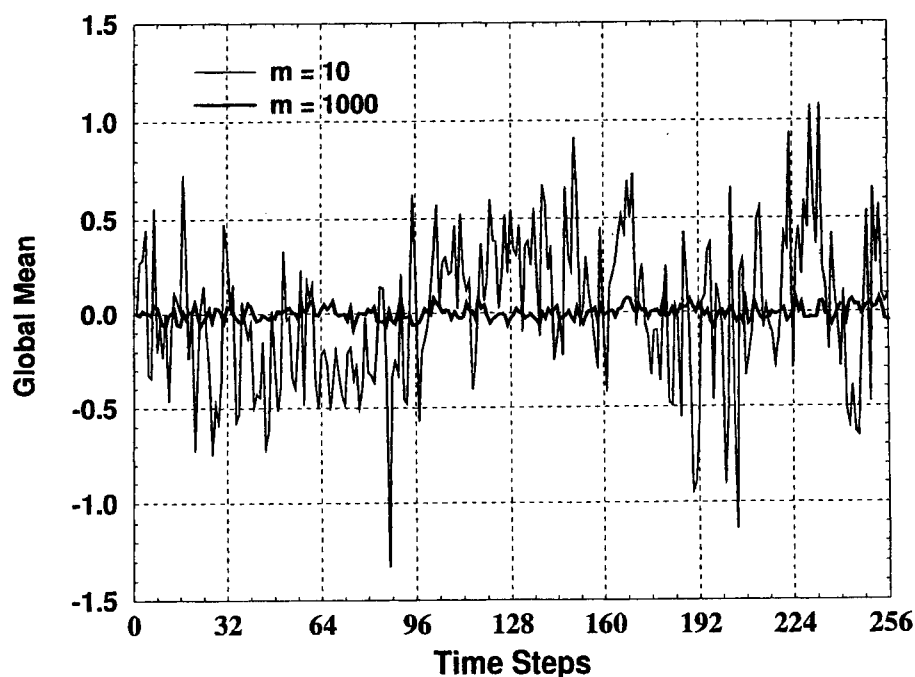
Figure 1. Simulated global mean value from m stations each one of them obeying the scaling law $F(t) \sim t^H$ with H=0.7 (see text for details).

References

1. Triantafyllou, G.N., R. Picard, R. and Tsonis, A.A., Exploiting geometric signatures to accurately derive properties of attractors. *Appl. Math. Lett* **7** (1994) pp. 19-24.
2. Tsonis, A.A. and J.B. Elsner, J.B., Testing for scaling in natural forms and observables. *J. Stat. Phys.* **81** (1995) pp. 869-880.
3. Viswanathan, G.M., Afanasyev, V., Buldyrev, S.V., Murphy, E.V., Prince, P.A. and Stanley, H.E., Lévy flight search patterns of wandering albatrosses. *Nature* **381** (1996) pp. 413-415.
4. Feder, J., Fractals (Plenum, New York, 1988).
5. Osborne, A.R., and Provenzale, A., Finite correlation dimension for stochastic systems with power-law spectra. *Physica D* **35**, (1989) pp. 357-381.
6. Tsonis, A.A., Chaos: From Theory to Applications , (Plenum, New York, 1992).
7. Tsonis, A.A., Roebber P.J. and J. B. Elsner, J.B., A characteristic time scale in the global temperature record. *Geophys. Res. Lett.*, **25** (1998) pp. 2821-2823.
8. Koscielny-Bunde, E., Bunde, A., Havlin, S., Roman, H.E., Goldreich Y. and Schellnhuber H-J, Indication of a universal persistence law governing atmospheric variability. *Phys. Rev. Lett.* **81** (1998) pp. 729-732.
9. Tsonis, A.A., Roebber, P.J. And Elsner, J.B., Long-range correlations in the extratropical atmospheric circulation: origins and implications. *J. Climate*, **12** (1999) pp. 1534-1541.

MULTISPECTRAL BACKSCATTERING : A FRACTAL-STRUCTURE PROBE

ROBERT BOTET[†] AND PASCAL RANNOU[‡]

[†] *Laboratoire de Physique des Solides - CNRS, Bâtiment 510, Université Paris-Sud*
Centre d'Orsay, F-91405 Orsay, France
Tel : (33/0) 1-69 15 69 25, Fax : (33/0) 1-69 15 60 86,
E-mail : botet@lps.u-psud.fr
and
[‡] *Service d'Aéronomie, Tour 15, Boîte 102, Université Paris 6,*
4 Place Jussieu, 75252 Paris Cedex 05, France
Tel : (33/0) 1-44 27 49 70, Fax : (33/0) 1-44 27 37 76,
E-mail : pra@ccr.jussieu.fr,

We present numerical results for backscattering of electromagnetic waves by fractal aggregates of fractal dimension ≤ 2. Single-scattering analytical result shows that Fourier transform of the density-density mass correlation function of the scatterer is the main contribution of the backscattering cross-section wavelength variation, when wavelength is in the fractal range of the scattering object. This allows to conjecture that multispectral backscattering analysis could be used to measure the geometrical fractal dimension of the scatterer. Illustrative results are given by using several recently proposed numerical scattering codes accounting for multiple scattering. Applications to lidar experiments in the visible range are briefly discussed.

1 Fractal Aggregates

Aerosol particles commonly consist of disordered clusters of nanometer-size grains. To give a few examples, these are soot particles from diesel engines [1], smoke obscurants [2] or haze particles in planetary atmospheres [3]. They correspond to a variety of standard two-step coalescence scenario : in a first regime, high-temperature nucleation creates compact small grains, then low-temperature aggregation forms disordered large clusters in which the grains, previously created, keep their individuality. Only the local junctions between neighbouring grains can be structurally modified by local melting, forming necks. These small local transformations will be neglected in the following, and all the grains will be considered as identical homogeneous spheres. All what will be written below for large fractal aerosol [4] clusters can be translated to disordered fractal colloidal aggregates as well, although the dispersing fluid medium is different, but the cluster structures are basically the same [5].

In dense atmosphere, two aggregation models are relevant [6] : the irreversible Brownian Cluster-Cluster Aggregation model (BrCCA) and the Reaction-limited Cluster-Cluster Aggregation model (RCCA). The first one (BrCCA) corresponds to the experimental conditions where sticking between grains is so strong that aggregation is irreversible and rapid [7] : sticking is permanent once formed. On the other hand, the latter model (RCCA) is the case where sticking is so weak that only
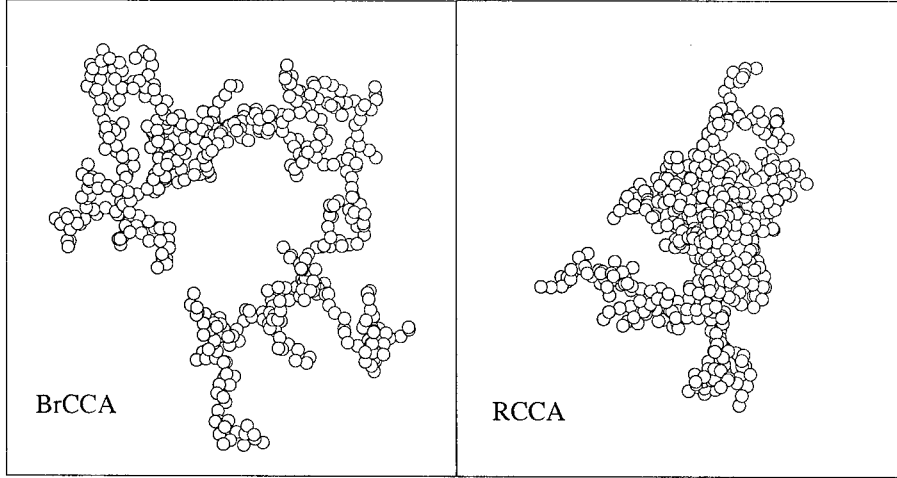
Figure 1. Fractal clusters of 512 identical particles generated respectively with the Brownian CCA model ($D_f = 1.75$), and with the Reaction-limited CCA model ($D_f = 2$)

the strongest structures are permanently formed [8] . In this slow aggregation, the clusters are more compact than for the BrCCA case. Both models lead to fractal structures, and the fractal dimensions are, respectively, $D_f = 1.75$ (for BrCCA) and $D_f = 2$ (for RCCA). Fig. 1 gives typical pictures of 3-dimensional clusters numerically built with such models.

Optical response of such structures carries informations about the fractal correlations. The most widely known is the so-called q^{-D_f}-law [9] valid for small-angle neutron- or X-scattering. This gives direct access to the fractal dimension of the clusters [10]. This law is derived in the case where scattering is so weak that multiple scattering can be neglected. It should be corrected when multiple scattering is expected to play some role - as when the refractive index is significantly different from 1, and fractal dimension larger than 2 -. Note that the value $D_f = 2$ is the expected threshold, some logarithmic corrections may arise in this case.

Properties of the scattering pattern of electromagnetic wave by such aggregate are the direct consequence of interferences between waves with different phases due to the spatial distribution of the grains. This distribution can be characterized by the distance-distribution function representing the density-density correlation function in the case of aggregates of N identical spheres [11] :

$$P(\vec{r}) = \left\langle \frac{1}{N(N-1)} \sum_{i \neq j} \delta(\vec{r} - \vec{r}_{ij}) \right\rangle \quad ,$$

where the summation runs over all pairs of distinct grains. The δ is the Dirac-distribution, and average is here over a large set of independent clusters of the same fractal dimension and the same number N of grains. This function $P(\vec{r})$, after orientation averaging, becomes a function of the variable r only. It can be

written [12] as :

$$P(r) = \frac{1}{A_{D_f} R_g^3} \left(\frac{r}{R_g} \right)^{D_f - 3} f_c \left(\frac{r}{R_g} \right) \quad , \tag{1}$$

with a positive cut-off function f_c. R_g is the radius of gyration, given by

$$R_g^2 = \frac{1}{2} \int_0^\infty r^2 P(r) 4\pi r^2 dr \quad .$$

The coefficient A_{D_f} in Eq.(1) insures proper normalization of the distance-distribution function. For simple applications, the cut-off function f_c is approximated by an exponentially decreasing function, such that [9] :

$$P(r) = \frac{1}{A_{D_f} R_g^{D_f}} r^{D_f - 3} \exp[-\frac{r}{R_g} \sqrt{\frac{D_f(D_f + 1)}{2}}] \quad . \tag{2}$$

This approximation is not valid in general and a much more precise form is : $f_c(z) \sim \exp(-z_f^D / 2)$ as it has been shown in [12]. The latter is useful for numerical computation, but analytic calculations can often be handled only for the exponential cut-off.

2 Electromagnetic Scattering Codes

First exact results for scattering of electromagnetic field on some large objects (i.e., of size comparable to the wavelength) were obtained in the classical Mie's work [13] for spherical, homogeneous scatterers irradiated by monochromatic plane wave. The solution is then given as a series of vector spherical functions which (usually) converges. There are several natural ways to extend this result to more complicated cases. One of them is to consider a rigid assembly of such identical balls, and to wonder if one can express the scattered field knowing the exact response of any of these balls. This is the T-matrix formalism which is briefly reviewed in the following section. Another way is to consider a non-homogeneous ball, where the refractive index depends continuously on the distance to the center, for example according to the averaged density of the cluster. This method is not well known and is discussed in more details below in Sec.D.

2.1 Orientation-Averaged T-Matrix Code

In principle, for the case of rigid aggregate of homogeneous spheres, the cluster T-matrix method is the exact solution of Maxwell equations with appropriate boundary conditions [14]. Its basis is just the linearity of Maxwell equations and boundary conditions. Writing the scattered electric field in terms of usual expansion :

$$\vec{E}_{sca}(j) = \sum_{n=1}^\infty \sum_{m=-n}^n iE_{mn}[a_{mn}^j \vec{N}_{mn}^{(3)}(j) + b_{mn}^j \vec{M}_{mn}^{(3)}(j)] \quad , \tag{3}$$

with the standard notations for the spherical harmonics functions [15], the method consists then in developing all these fields over a unique ensemble of vector spherical functions, by systematic use of the addition theorem which expresses spherical

harmonic functions referring to the center of some grain into series of vector spherical functions with the origin on a reference point. All the fields, included the incident one, being expressed over the same basis of vectors, one can write self-consistently that the local field on particle j is the sum of the incident field plus all fields scattered by other particles, and obtain a set of coupled linear equations in all the expansion coefficients (a_{mn}^j, b_{mn}^j). Truncating the order of these equations at a reasonable level of accuracy, the problem reduces to an inversion of a large matrix equation. Moreover, a very efficient computational method has recently been proposed, where orientation averaging is performed analytically [17]. Since fractal clusters are essentially non-spherical, *except in a statistical meaning*, this scheme provides considerable help in the practical computation of the average scattering coefficients.

2.2   Mean-Field Mie-Code

The T-matrix equations can be substantially simplified in a mean-field approach where all the grains are supposed to scatter the *same* electromagnetic field. This has been done previously for the dipolar scattered fields [18], but this was recently extended to the Mie scattering [19], that is, with scattered fields of the form (3). In this case we have not to consider self-consistent equations between N scattered fields, but just one self-consistent equation with one (mean-field) scattered field. Inversion of a small matrix equation is needed and usually can be performed quickly on any small computer. It is difficult to judge the correctness of the mean-field hypothesis, but numerically it leads to one of the best, now available approximations for self-similar aggregates of identical grains [16]. Qualitatively, great inhomogeneities of the local fields should be avoided and this means that the system is far from any optical resonance [20]. Moreover, the geometrical structure itself should be as homogeneous as possible. This means for fractal aggregates that we have to deal only with strictly self-similar objects, like the clusters generated by the models discussed here (BrCCA and RCCA).

2.3   Discrete-Dipole Approximation

This method, known as DDA numerical method [21] , consists in dividing the scatterer into identical pieces small enough to be individually considered as electromagnetic dipoles, but large enough in order that the number of such sub-units be not too large. All these dipoles interact, and since the field radiated by each dipole is analytically simple, the method is exact, as far as space discretisation is fine enough to insure the dipolar representation be correct. Below, we used this method with one dipole per grain. This is correct if the wavelength is much larger than the radius of one grain, but as soon as the size parameter becomes of order unity, we should take several dipoles per grain, but the computations become rapidly important from both CPU time and memory points of view.

2.4 Fractally-Coated Code

This section is about a new approximate computation of the scattered properties for fractal clusters. Instead of taking the exact solution by the T-matrix method, then averaging over all possible orientations, one can decide first to average the geometric structure, yielding radially decreasing density of matter, and then to compute the scattering properties as if it was for a ball of same spherical density. More precisely, consider fractal distributions of matter with fractal dimension D_f. Suppose that these are fractal aggregates of balls of radius a. If they are statistically radially symmetric, one can identify the center of the ball closest to the center of gravity as the center of coordinates and write the average density as :

$$\begin{aligned}
\rho(r) &= 1 & \text{if} \quad & r \leq a & (4) \\
\rho(r) &= \left(\frac{r}{a}\right)^{D_f - 3} & \text{if} \quad & a < r < R \\
\rho(r) &= 0 & \text{if} \quad & r > R \quad,
\end{aligned}$$

where R is a typical radius of the aggregate. The constraint on the total mass of the aggregate gives explicitly the value of R since we must have

$$M(\infty) = \int_0^R \left(\frac{r}{a}\right)^{D_f - 3} 4\pi r^2 dr \quad,$$

equal to the mass of the aggregate, namely, $M = 4\pi a^3 N/3$. This means that

$$R/a = (D_f N/3)^{1/D_f} \quad.$$

With these results in hands one can compute the electromagnetic field scattered by a ball of radial refractive index $n_{av}(r)$ given by the Maxwell-Garnett rule [15]

$$n_{av}^2(r) = \frac{n^2(1 + 2\rho(r)) + 2(1 - \rho(r))}{n^2(1 - \rho(r)) + 2 + \rho(r)} \quad,$$

with n the refractive index for the grain. This comes from the multilayered sphere recursive equations [22] which are both stable and accurate.

2.5 Numerical Comparison Between Codes

All four methods have been implemented into numerical codes for fractal aggregates. Firstly we compare all of them, the T-matrix method being considered as giving the exact results. We present in this article a small part of this task, more systematic comparisons are in preparation [23]. We have decided to show here some comparisons for several representative values of refractive index, namely, $n = 1.5$ (non absorbing), $n = 1.5 + i * 10^{-4}$, $n = 1.5 + i * 10^{-2}$ and $n = 1.5 + i$ (strongly absorbing). Just one size is discussed, $N = 512$ - for which the fractal features are known to be well developed -, and the grain-size parameter $x = ka = 2\pi a/\lambda$ is between 10^{-4} and 1. For these parameters, the T-matrix as well as the DDA methods cannot go to values of x much larger than 1. The extinction and backscattering cross-sections are shown for the four computations in Figs. 2, and 3.
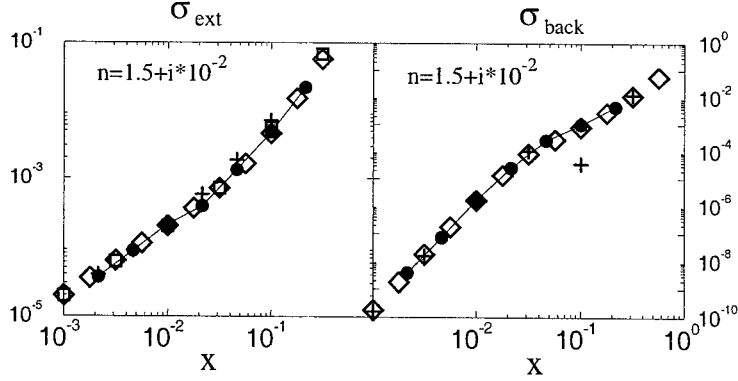
Figure 2. Numerical comparisons between the four methods described in the text, for $N = 512$ RCCA clusters of fractal dimension $D_f = 2$. The refractive index is $n = 1.5 + i * 10^{-2}$, and the grain-size parameter x varies between 10^{-3} and 1. Double-logarithmic scale, the cross-sections have all been divided by $N\pi a^2$ for normalisation. Symbols are : full circles (and line) are for T-matrix, diamonds for Mean-Field Mie, crosses for DDA, squares for fractally-coated methods, respectively.
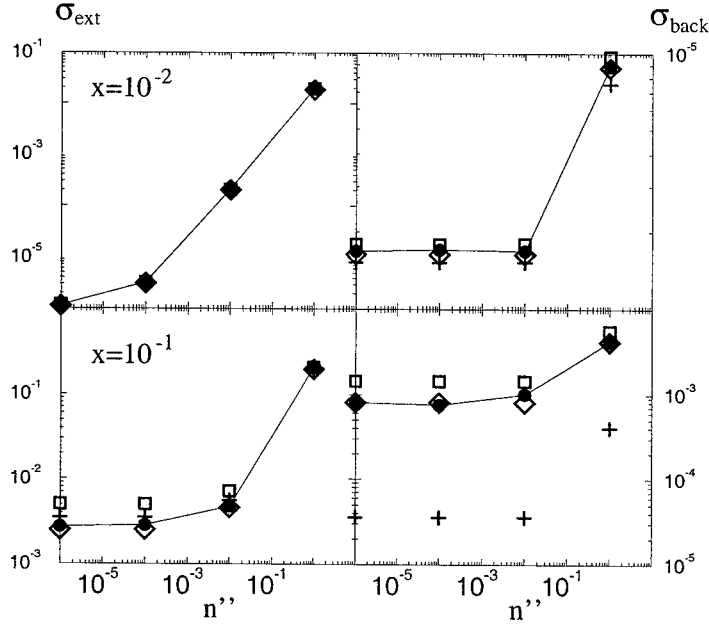
Figure 3. Numerical comparisons between the four methods described in the text, for $N = 512$ RCCA clusters of fractal dimension $D_f = 2$. The refractive index is $n = 1.5 + i * n"$, with different values of $n"$ between 0 and 1. Note that the data for $n" = 0$ (i.e. for the non-absorbing index : $n = 1.5$) has been shifted to $n" = 10^{-6}$ to show them on the same figure. Double-logarithmic scale, the cross-sections have all been divided by $N\pi a^2$ for normalisation. Symbols are : full circles (and line) are for T-matrix, diamonds for Mean-Field Mie, crosses for DDA, squares for fractally-coated computations respectively.

The results are quantitatively comparable even if the fractally-coated results are indeed less precise, leading to slight overestimation of the cross-sections when x becomes large. The DDA code may give wrong under-estimate values of the backscattering cross-sections, due to lack of numerical stability, as shown in the $x = 10^{-1}$ case. These results reinforce the conclusions that the Mean-Field Mie method is the most reliable approximation up to the largest values of x attainable by T-matrix code with our computers.

3   Backscattering Coefficient

The backscattering coefficient needs particular comment. This coefficient is important in some applications like lidar experiments, basically because in such experiments, one light beam is emitted and the backscattered light is analysed [24]. This kind of experiments can be handled with far scattering sources because of this simple geometry, and gives access to precise informations about remote aerosol scatterers [25]. Furthermore, backscattering is also phenomenologically known to be important to distinguish between various forms of scatterers [26].

3.1   Single-Scattering Theory

If we suppose single-scattering to be the dominant feature, and the cut-off function to be exponential, then the orientation-averaged intensity I_s of the light scattered at angle θ by fractal aggregate of size N and fractal dimension D_f, is just [9]

$$I_{sca}(N) = N I_{sca}(1) \left[1 + (N-1) \frac{\sin[(D_f - 1)\tan^{-1}(q\xi)]}{(D_f - 1)q\xi} \frac{1}{(1 + q^2\xi^2)^{(D_f-1)/2}} \right] \quad (5)$$

with the modulus of the scattering vector, $q = 2k\sin(\theta/2)$, and $\xi = R_g\sqrt{2/D_f(D_f + 1)}$. This formula allows to recover properly the Guinier-regime for $qa \ll 1$ and the fractal q^{-D_f}-law for $1/\xi \ll q \ll 1/a$. Moreover, for extinction, we are interested in $q = 0$ since we look at the $\theta = 0$ case, and so

$$\sigma_{ext}(N) = N\sigma_{ext}(1) \quad ,$$

as a consequence of forward coherence. More precise forms of extinction cross-section can be found in [18].

The case is totally different for backscattering since in this case, $q = 2k$, and the waves scattered by the grains do not have the same phase. More precisely :

$$\sigma_{back}(N) = N\sigma_{back}(1) \left[1 + (N-1) \frac{\sin[(D_f - 1)\tan^{-1}(2k\xi)]}{(D_f - 1)2k\xi} \frac{1}{(1 + 4k^2\xi^2)^{(D_f-1)/2}} \right] \quad ,$$

as an immediate consequence of (5). In particular, in the case of clusters large compared to the wavelength ($k\xi \gg 1$), the preceding formula approximates as

$$\sigma_{back}(N) \simeq N\sigma_{back}(1) \left[1 + \frac{B_{D_f}}{(ka)^{D_f}} \right] \quad ,$$

with finite numerical coefficient B_{D_f} depending only on the fractal dimension of the scatterer. This means that when ka is small enough, the ratio $\sigma_{back}(N)/N\sigma_{back}(1)$

is essentially proportional to $(ka)^{-D_f}$, or equivalently

$$\ln\left(\frac{\sigma_{back}(N)}{N\sigma_{back}(1)}\right) \sim D_f \ln(\lambda) \quad \text{for} \quad a < \lambda < R_g. \tag{6}$$

When $\lambda \ll a$, then the ratio $\sigma_{back}(N)/N\sigma_{back}(1)$ is close to 1 and its logarithm vanishes. For $\lambda \gg R_g$, the phase difference between the grains can be neglected and we recover the same coherence peak as for forward scattering [27] $\sigma_{back}(N)/N\sigma_{back}(1) \simeq N$ independently of the wavelength. Relation (6) provides a possible measure of the fractal dimension D_f of the scatterers analysed at different wavelengths. This is in principle a natural way to examine a fractal aggregate since the basic property of such an object is self-similarity. We just claim that analysing a fractal with different wavelengths correspond to investigating its structure with different length scales, and the result (6) is an illustration of this fact. A particular case of this idea has previously been proposed [28] with application to fractal Sierpinski gaskets .

3.2 Multiple-Scattering Results

So, in principle, relation (6) should give access to the fractal dimension of the clusters if one can measure the backscattering cross-section at different wavelengths in between the typical size of the grains and the typical radius of the aggregates. The result is in principle strictly valid only for the single-scattering approximation. One can wonder then if it holds also when multiple-scattering is no more negligible. In fact, we can qualitatively argue that for fractal dimensions $D_f \leq 2$, multiple scattering is expected to be small compared to single scattering. So, in this case, the signature of the fractal structure should be seen anyway, even if a bit modified. This should be checked on numerical simulations taking the multiple scattering into account.

But first, we have to be careful about the fact that the wavelength dependence of the refractive index of grain material is not generally cancelled by such simple procedure as taking the ratio $\sigma_{back}(N)/N\sigma_{back}(1)$, when multiple scattering is present. Focusing on the signature of fractal structure, we suppose here that correction is done in such a way that the refractive index of the grains can be considered as constant over the range of wavelengths. Usually, this can be done when the wavelength dependence of the refractive index of the material and the typical size of the grains are known.

Since we have shown before that the Mean-Field Mie code gives results very close to the exact T-matrix computation inside the range of analysis, we shall discuss the results of this method (Mean-Field Mie code) which can easily be used for very large values of the grain-size parameter.

Fig. 4 gives the numerical results for $N = 512$, and the two kinds of fractal aggregates : BrCCA with $D_f = 1.75$ and RCCA with $D_f = 2$, for the refractive index $n = 1.5 + i*10^{-2}$. The single-scattering result (6) is well recovered over about one decade in both cases. This range is the expected one since for these models and $N = 512$, the values of ξ/a are about 20 and 14 for BrCCA and RCCA respectively. The left-hand plateau for the long wavelengths is the Rayleigh coherent regime
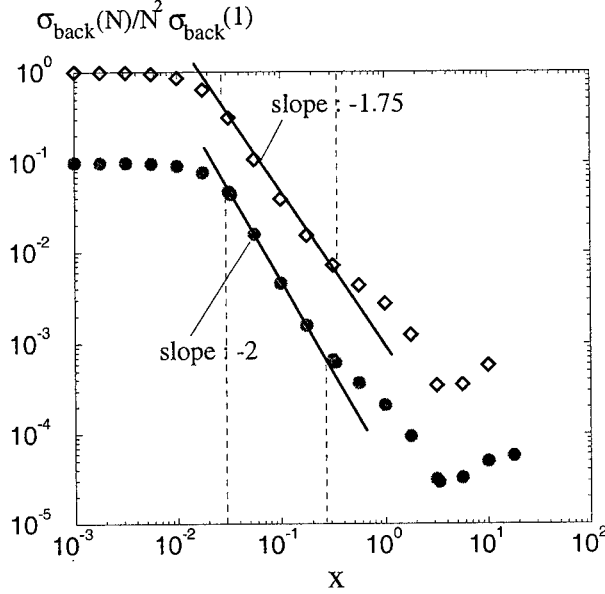
Figure 4. Sketch of the ratio $\sigma_{back}(N)/N^2\sigma_{back}(1)$ *vs* the grain-size parameter x for BrCCA (diamonds, fractal dimension = 1.75) and RCCA (circles, fractal dimension = 2), calculated with the Mean-Field Mie method. Double-logarithmic plot. The RCCA data have been divided by a factor 10 for visual separation of the two cases. The D_f-slope - derived in the single-scattering theory -, is well recovered between the dashed lines.

and in this case we expect : $\sigma_{back}(N) \simeq N^2\sigma_{back}(1)$. On the other side, for the small wavelengths, one begins to see oscillating behaviour on the curves, before the possible limiting value $\sim 1/N$. These results show clearly that the single-scattering result (6) is very robust in these cases and dominates the leading behaviour even if multiple scattering is taken into account.

4   Conclusions

We briefly reviewed in this work four theoretical methods allowing to compute either exactly or with fine accuracy the optical cross-sections of fractal aggregates. We used the corresponding numerical codes to compare their results for two sets of typical fractal clusters of fractal dimension $D_f = 1.75$ and $D_f = 2$, relevant in aerosol and colloid physics. For reasonable values of the refractive index of the material, the Mean-Field Mie method is seen to give the most accurate and reliable results, as compared to the exact T-matrix method, which is limited by the large amount of computations when the grain-size parameter becomes large.

As an application, we have studied the wavelength dependence of the backscattering cross-section for wavelengths between the size of the grains and the radius of the aggregates. It has been shown that such analysis gives access to a very particular signature of the fractal structure, and could in principle be used to estimate the

38

fractal dimension of the scatterers by lidar experiments, since lidar signal depends essentially of the backscattering and the extinction cross-sections of the scattering aerosols. This should give a new alternative to measure experimentally the fractal dimension of fractal aggregates by electromagnetic radiation scattering.

Note : The numerical scattering codes for fractal aggregates are available upon request from *pra@ccr.jussieu.fr* for the Mean-Field Mie and DDA methods, and from *botet@lps.u-psud.fr* for the T-matrix and fractally-coated methods. The basic T-matrix code may be found at *http://www.giss.nasa.gov/ crmim/*, and original DDA code comes from *http://www.astro.princeton.edu/ draine/scattering.html*.

References

1. Sorensen C., Cai J., Lu N., Appl. Opt. **31** (1992) p 9547-6557.
   Mengüç M., Manckavasagam S., Appl. Opt. **36** (1997) p 1337-1351.
2. Dziedzinl F., Botet R., J. Phys. II **1** (1991) p 343-352.
3. West R.A., Smith P.H., Icarus, **90** (1991) p 330-333.
   Rannou P., Cabane M., Chassefière E., Botet R., McKay C.P., Icarus, **111** (1995) p 355-372.
4. Forrest S., Witten T., J. Phys. A, **12** (1979) p L109-L117.
5. Weitz D., Lin M., Sandroff C., Surf. Sci. **158** (1985) p 147-164.
6. Jullien R., Botet R., *Aggregation and Fractal Aggregates*, World Scientific, Singapore 1987.
7. Meakin P., Phys. Rev. Lett., **51** (1983) p 1119-1122.
   Kolb M., Botet R., Jullien R., Phys. Rev. Lett., **51** (1983) p 1123-1126.
8. Kolb M., Jullien R., J. Physique Lett. **45** (1984) p L977-981.
   Brown W.D., Ball R., J. Phys. A **18** (1985) p L517-L521.
9. Teixeira J., in *On Growth and Form*, Proceedings of the NATO Advanced Study Institute, Martinus Nijhoff, Dordrecht Netherthelands 1986 p 145-162.
10. Schaefer D.W., Martin J.E., Wiltzius P., Cannell D.S., Phys. Rev. Lett., **52** (1984) p 2371-2374.
11. Meakin P., J. Phys. A **18** (1985) p L661-L666.
12. Botet R., Rannou P., Cabane M., J. Phys. A **28** (1995) p 297-316.
13. Mie G., Ann. Phys. **25** (1908) p 377-452.
14. Mackowski D.W., Proc. R. Soc. London Ser.A **433** (1991) p 599-614.
    Xu Y.-L., Appl. Opt. **34** (1995) p 4573-4588.
15. Bohren C.F., Huffman D.R., *Absorption and Scattering of Light by Small Particles*, John Wiley & Sons, New-York 1983.
16. Rannou P., Cabane M., Botet R., Chassefière E., J. Geophys. Research, **102** (1997) p 10997-11013.
17. Khlebtsov N.G., Appl. Opt., **31** (1992) p 5359-5365.
    Mackowski D.W., Mishchenko M.I., J. Opt. Soc. Am. **13** (1996) p 2266-2278.

18. Berry M.V., Percival I.C., Optica Acta **33** (1986) p 577-591.

19. Botet R., Rannou P., Cabane M., Appl. Opt. **36** (1997) p 8791-8797.

20. Shalaev V.M., Botet R., Tsai D.P., Kovacs J., Moskovits M., Physica A **207** (1994) p 197-207.

21. Purcell E.M., Pennypacker, Astrophys. J. **186** (1973) p 705-514.
    Draine B.T., Flatau P.J., J. Opt. Soc. Am A **11** (1994) p 1491-1499.

22. Kerker M., *The Scattering of Light and Other Electromagnetic Radiations*, Acad. San Diego, California 1969.
    Wu Z.S., Wang Y.P., Radio Science **26** (1991) p 1393-1401.

23. Botet R., Rannou P., in preparation.

24. Klett J.D., Appl. Opt., **20** (1981) p 211-220.
    Kasparian J., Frejafon E., Rambaldi P., Yu J., Vezin B., Wolf J.-P., Atm. Env. **32** (1998) p 2957-2967.

25. Frejafon E., Kasparian J., Rambaldi P., Vezin B., Yu J., Wolf J.-P., Appl.Opt. **37** (1998) p 2231-2237.

26. Barber P.W., Massoudi H., Aerosol Sci. Technol. **1** (1982) p 303-310.

27. Asakura T., Ishii K., Iwai T., Uozumi J., Appl. Opt. **37** (1998) p 5014-5018.

28. Lakhtakia, Messier R., Varadan V.V., Varadan V.K., J. Phys. A **20** (1987) p 1617-1619.

SMALL-ANGLE MULTIPLE SCATTERING ON A FRACTAL SYSTEM OF POINT SCATTERERS

V.V.UCHAIKIN

Ulyanovsk State University
Institute for Theoretical Physics
42 L.Tolstoy Str., 432700 Ulyanovsk, Russia
E-mail: uchaikin@sv.uven.ru

Multiple scattering of classical particles on a system of point scatterers distributed in space in a fractal fashion is considered in the small-angle approximation. The asymptotic regime of this process is described by a fractional differential equation generalizing the ordinary diffusion in the angle space equation. The solution of the problem is presented both in terms of stable distributions and in terms of Fox-functions. Main features of the obtained solutions are discussed.

1 Introduction

Small-angle x-ray and neutron scattering experiments are efficient tools for studying the structure of condensed matter. Measurements of the light or cosmic rays from distant sources play the same role in studying the large-scale structure of the Universe. From mathematical point of view, the problems belong to the class of inverse problems solved in terms of the multiple scattering theory.

The ordinary multiple scattering theory assumes that the random spatial distribution of scatterers is a homogeneous Poisson ensemble, i.e. different scatterers are placed independently of each other. That is just what guarantees the exponential character of the free path distribution and leads to the Boltzmann master equation. In the region of large depths and small angles the equation takes the form of the diffusion (in angle space) equation called the Fokker-Planck equation. Its solution is nothing but the two-dimensional (in angle space) Gauss distribution called also the Fermi distribution. However, above-mentioned assumption becomes invalid for fractal media characterized by long-range correlations of the power type.

One approach to solution of the problem is developed by S.Maleyev [1]. He has considered a medium with fractal pores distributed uniformly in space. As a result the superdiffusion behaviour of the angular distribution has been found. Another approach is developed in our previous work [2], where the fractal medium is considered as a set of identical point scatterers distributed in space in fractal manner by means of the Lévy flight procedure (see for details [3,4,5]). This model reveals the subdiffusion behaviour of the angle distribution.

The present work combines both of these ideas. We consider here a set of fractal clusters distributed in space in a fractal manner. As a result, we obtain a two-parametrical family of angular distributions. The family includes the Fermi distribution, the subdiffusion and superdiffusion distributions as particular cases. They obey fractional differential equations and are presented both in terms of stable distributions and in terms of Fox-functions.

2 Basic postulates of the small-angle scattering theory

Let a particle move from the origin of coordinates along the x-axis. After the scattering, the particle will be characterized by the vector angle $\vec{\Theta}$ of deviation from the initial direction $\vec{\Omega}_0$ of the motion

$$\vec{\Theta} = \vec{\Omega} - \vec{\Omega}_0.$$

The following postulates form the well-known basis of the small-angle approximation in the scattering theory.

i) The particle is on the x-axis during all its motion even though the (small) vector $\vec{\Theta}$ differs from zero.

ii) Vector $\vec{\Theta}$ stays constant between points of scatterings $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$ and undergoes random jumps $\vec{\Theta}_1, \vec{\Theta}_2, \vec{\Theta}_3, \ldots$ at these points, so

$$\vec{\Theta}(x) = \sum_{i=1}^{N(x)} \vec{\Theta}_i$$

where $N(x)$ is the random number of scatterings on the segment $[0, x]$.

iii) The random variables $X_1, X_2, X_3, \ldots$ are independent of each other and have the same distribution density $q(x)$.

iv) The random variables $\vec{\Theta}_1, \vec{\Theta}_2, \vec{\Theta}_3, \ldots$ are independent of each other and of $X_1, X_2, X_3, \ldots$ and identically distributed with the density $\sigma(\vec{\theta})$.

v) The distribution $\sigma(\vec{\theta})$ is concentrated in such a small region near zero that the vectors $\vec{\Theta}_i$ and $\vec{\Theta}(x)$ may be considered as two-dimensional vectors and integration with respect to $\vec{\theta}$ may be extended to the whole plane $\mathbf{R}^2$.

On these assumptions, the angular distribution density $p(\vec{\theta}, x)$ of the particle passed the path x is given by the sum

$$p(\vec{\theta}, x) = \delta(\vec{\theta})Q(x) + \sigma(\vec{\theta}) \int_0^x Q(x - x')q(x')dx' +$$

$$\sigma * \sigma(\vec{\theta}) \int_0^x Q(x - x')q * q(x')dx' +$$

$$\sigma * \sigma * \sigma(\vec{\theta}) \int_0^x Q(x - x')q * q * q(x')dx' + \ldots \tag{1}$$

Here $*$ means convolution of distributions

$$q * q(x) \equiv \int_0^x q(x - x')q(x')dx',$$

$$\sigma * \sigma(\vec{\theta}) \equiv \int_{\mathbf{R}^2} \sigma(\vec{\theta} - \vec{\theta}')\sigma(\vec{\theta}')d\vec{\theta}',$$

and

$$Q(x) \equiv \int\limits_x^\infty q(x')dx'.$$

Represent (1) in the form

$$p(\vec{\theta}, x) = \int\limits_0^x Q(x - x')f(\vec{\theta}, x')dx'. \tag{2}$$

The function $f(\vec{\theta}, x)$ is the density of collisions and can be expanded in the collision number series

$$f(\vec{\theta}, x) = \delta(\vec{\theta})\delta(x) + \sigma(\vec{\theta})q(x) +$$

$$+\sigma * \sigma(\vec{\theta})q * q(x) + \sigma * \sigma * \sigma(\vec{\theta})q * q * q(x) + \ldots =$$

$$= \sum_{n=0}^\infty \sigma^{(n)}(\vec{\theta})q^{(n)}(x), \tag{3}$$

where

$$\sigma^{(0)}(\vec{\theta}) = \delta(\vec{\theta}), \quad \sigma^{(1)}(\vec{\theta}) = \sigma(\vec{\theta}), \quad \sigma^{(n)}(\vec{\theta}) = \underbrace{\sigma * \ldots \sigma(\vec{\theta})}_{n}, \quad n \geq 1$$

and so are $q^{(n)}(x)$. It is easy to see that the distribution (3) is a solution of the master equation

$$f(\vec{\theta}, x) = \delta(\vec{\theta})\delta(x) + \int\limits_0^x dx' q(x - x') \int\limits_{\mathbf{R}^2} d\vec{\theta}' \sigma(\vec{\theta} - \vec{\theta}')f(\vec{\theta}', x'). \tag{4}$$

Formula (2) together with master equation (4) gives a complete description of the problem.

3 The scattering theory for a homogeneous medium

The ordinary theory of multiple scattering appears as a result of the addition of two further postulates to those mentioned above:

vi) The mean square angle of a single scattering has a finite value

$$\langle \Theta^2 \rangle = \int\limits_{\mathbf{R}^2} \sigma(\vec{\theta})\theta^2 d\vec{\theta} = 2\pi \int\limits_0^\infty \sigma(\theta)\theta^3 d\theta < \infty.$$

vii) The random variables $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$ form the Poisson ensemble, i.e. the increments $X_1, X_2, X_3, \ldots$ are distributed according to the exponential law

$$q(x) = \mu e^{-\mu x}, \quad \mu > 0. \tag{5}$$

44

Inserting (5) into (2) and (4) we obtain two expressions

$$p(\vec{\theta}, x) = e^{-\mu x} \int_0^x e^{\mu x'} f(\vec{\theta}, x') dx',$$

$$f(\vec{\theta}, x) = \delta(\vec{\theta})\delta(x) + \int_{R^2} d\vec{\theta}' \sigma(\vec{\theta} - \vec{\theta}') \mu e^{-\mu x} \int_0^x e^{\mu x'} f(\vec{\theta}', x') dx'$$

which after simple transformations take a form of the integro-differential kinetic equation (the Boltzmann equation)

$$\frac{\partial p(\vec{\theta}, x)}{\partial x} = -\mu p(\vec{\theta}, x) + \mu \int_{R^2} d\vec{\theta}' \sigma(\vec{\theta} - \vec{\theta}') p(\vec{\theta}', x)$$

with the initial condition

$$p(\vec{\theta}, 0) = \delta(\vec{\theta}).$$

In the large x-asymptotics, the equation reduces to the ordinary diffusion equation

$$\frac{\partial p(\vec{\theta}, x)}{\partial x} = D\nabla_\theta^2 p(\vec{\theta}, x) \tag{6}$$

with $D = \mu \langle \Theta^2 \rangle / 2$. It is supposed here that $\sigma(\theta)$ is an axially symmetric function depending on $|\vec{\theta}| = \theta$.

4 The scattering theory for a fractal medium

A passage from a homogeneous medium to a fractal one is performed by the replacement of postulates vi) and vii) with the following:

vi)* The angular distribution of particles scattered on a single fractal cluster is characterized by the power law [1]

$$\int_{|\vec{\theta}'| > \theta} \sigma(\vec{\theta}') d\vec{\theta}' \sim A\theta^{-\alpha}, \quad \theta \to \infty, \quad 0 < \alpha < 2. \tag{7}$$

vii)* The free path distribution in a fractal path medium has a long power tail too [2]:

$$Q(x) \sim Bx^{-\beta}, \quad x \to \infty, \quad 0 < \beta < 1. \tag{8}$$

Note that the mean square angle in the first case and mean free path in the second one are infinite.

The information given by (7) and (8) is enough to find the asymptotic behaviour of the distribution $p(\theta, x)$ at large distances x. To do this we apply the Fourier-Laplace transform

$$p(\vec{k}, \lambda) = \int_{R^2} d\vec{\theta} \int_0^\infty dx e^{i\vec{k}\vec{\theta} - \lambda x} p(\vec{\theta}, x),$$

where $\vec{k}$ is a two-dimensional vector and $\vec{k}\vec{\theta}$ is the scalar product. Formulas (2) and (4) take the following form:

$$p(\vec{k}, \lambda) = [1 - q(\lambda)]f(\vec{k}, \lambda)/\lambda,$$

$$f(\vec{k}, \lambda) = 1 + q(\lambda)\sigma(\vec{k})f(\vec{k}, \lambda).$$

Combining them we get

$$p(\vec{k}, \lambda) = \frac{1 - q(\lambda)}{\lambda[1 - q(\lambda)\sigma(\vec{k})]}.$$

This is the Montroll-Weiss result for random walk with trapping in a two-dimensional space [6] (see, also [7] and [8]).

According to the Tauberian theorem [9], the relations (7) and (8) determine behaviour of transforms $\sigma(\vec{k})$ and $q(\lambda)$ at small arguments:

$$1 - \sigma(\vec{k}) \sim ak^\alpha, \quad k \to 0, \quad a = \frac{2^{-\alpha}\Gamma(1 - \alpha/2)}{\Gamma(1 + \alpha/2)}A,$$

$$1 - q(\lambda) \sim b\lambda^\beta, \quad \lambda \to 0, \quad b = \Gamma(1 - \beta)B.$$

In their turn, these expressions make it possible to find the asymptotic form for $p(\vec{k}, \lambda)$:

$$p(\vec{k}, \lambda) \sim p_{as}(\vec{k}, \lambda) = \frac{\lambda^{\beta-1}}{\lambda^\beta + Ck^\alpha}, \quad \lambda \to 0, \quad k \to 0 \tag{9}$$

with $C = a/b$, whence

$$p_{as}(\vec{\theta}, x) = \frac{1}{(2\pi)^3 i} \int\limits_L d\lambda \int\limits_{R^2} d\vec{k} \frac{\lambda^{\beta-1}}{\lambda^\beta + Ck^\alpha} e^{-i\vec{k}\vec{\theta} + \lambda x} =$$

$$= (Cx^\beta)^{-2/\alpha} \Psi_2^{(\alpha,\beta)}(\vec{\theta}(Cx^\beta)^{-1/\alpha}) \tag{10}$$

where

$$\psi_2^{(\alpha,\beta)}(\vec{\theta}) = \frac{1}{(2\pi)^3 i} \int\limits_{L'} d\lambda \int\limits_{R^2} d\vec{k} \frac{\lambda^{\beta-1}}{\lambda^\beta + k^\alpha} e^{-i\vec{k}\vec{\theta} + \lambda}. \tag{11}$$

On the other hand, Eq. (9) rewritten in the form

$$\lambda^\beta p_{as}(\vec{k}, \lambda) = -Ck^\alpha p_{as}(\vec{k}, \lambda) + \lambda^{\beta-1},$$

can be considered as the Fourier-Laplace transform of the fractional differential equation. In terms used in the book [10] it has the form:

$$D_{x0+}^\beta p_{as}(\vec{\theta}, x) = -C(-\Delta_\theta)^{\alpha/2} p_{as}(\vec{\theta}, x) + \frac{x^{-\beta}}{\Gamma(1 - \beta)}\delta(x) \tag{12}$$

with the initial condition

$$p_{as}(\vec{\theta}, 0) = \delta(\vec{\theta}).$$

When $\alpha \to 2$ and $\beta \to 1$ the equation reduces to the ordinary diffusion equation (6) with $D = C$.

5   Results

The integral (11) giving the solution of the equation (12) via (10) may be represented in two forms [11]. The first of them is

$$\Psi_2^{(\alpha,\beta)}(\vec{\theta}) = \int\limits_0^\infty g_2^{(\alpha)}(\vec{\theta}x^{\beta/\alpha})g_1^{(\beta)}(x)x^{2\beta/\alpha}dx \tag{13}$$

where $g_1^{(\beta)}(x)$ and $g_2^{(\alpha)}(\vec{\theta})$ are the one-dimensional one-sided stable density and the two-dimensional axially symmetrical stable density determined by their Laplace and Fourier transforms respectively:

$$\int\limits_0^\infty e^{-\lambda x}g_1^{(\beta)}(x)dx = e^{-\lambda^\beta}$$

and

$$\int\limits_{R^2} e^{i\vec{k}\vec{\theta}}g_2^{(\alpha)}(\vec{\theta})d\vec{\theta} = e^{-k^\alpha}.$$

Stable distributions are described in detail in [11,12].

The second form uses the H-functions – generalized hypergeometrical functions, also called the Fox functions [13],

$$H_{mn}^{pq}\left(z\left|\begin{array}{c}(a_1,\alpha_1)\ldots(a_p,\alpha_p)\\(b_1,\beta_1)\ldots(b_q,\beta_q)\end{array}\right.\right) = \sum_{j=1}^m\sum_{k=0}^\infty \frac{(-1)^k}{k!}c_{jk}z^{s_{jk}}/\beta_j,$$

where m, n, p, and q are integer numbers such that $0 \le n \le p$ and $1 \le m \le q$; α_i and β_j are positive,

$$s_{jk} = (b_j + k)/\beta_j$$

$$c_{jk} = \frac{\prod\limits_{l=1,l\neq j}^m \Gamma(b_l - \beta_l s_{jk}) \prod\limits_{l=1}^n \Gamma(1 - a_l + \alpha_l s_{jk})}{\prod\limits_{l=m+1}^q \Gamma(1 - b_l + \beta_l s_{jk}) \prod\limits_{l=n+1}^p \Gamma(a_l - \alpha_l s_{jk})}.$$

In terms of H-function, the angular function $\Psi_2^{(\alpha,\beta)}(\vec{\theta})$ looks as follows

$$\Psi_2^{(\alpha,\beta)}(\vec{\theta}) = \frac{\beta}{8\pi}\left(\frac{2}{\theta}\right)^{2+\alpha}H_{32}^{12}\left(\left(\frac{2}{\theta}\right)^\beta\left|\begin{array}{c}(-1,1/\alpha)(-\alpha/2,\beta/2)(1-\alpha/2,1/2)\\(0,1/\alpha)(-1,1/\alpha)\end{array}\right.\right), \quad \alpha <$$

$$\Psi_2^{(\alpha,\beta)}(\vec{\theta}) = \frac{1}{\pi(\alpha\theta)^2}H_{23}^{21}\left(\frac{\theta}{2}\left|\begin{array}{c}(1,1/\alpha)(1,\beta/\alpha)\\(1,1/\alpha)(1,1/2)(1,1/2)\end{array}\right.\right), \quad 1 \le \alpha \le 2$$

When $\alpha = 2$ and $\beta = 1$

$$g_1^{(1)}(x) = \delta(x - 1),$$

$$g_1^{(2)}(\vec{\theta}) = \frac{1}{4\pi} \exp\left\{-\theta^2/4\right\} \tag{14}$$

and the density (13) reduces to the ordinary result for a homogeneous medium

$$p_{as}(\vec{\theta}, x) = \frac{1}{4\pi dx} \exp\left\{-\theta^2/(4Cx)\right\}$$

with the diffusivity C. This is the two-dimensional Gauss distribution with the mean square angle

$$\langle \theta^2 \rangle = \int\limits_{\mathbf{R}^2} p_{as}(\vec{\theta}, x)\theta^2 d\vec{\theta} = 4Cx, \tag{15}$$

describing a normal diffusion.

When $\alpha < 2$ and $\beta = 1$ we obtain

$$p_{as}(\vec{\theta}, x) = (Cx)^{-2/\alpha} g_2^{(\alpha)}(\vec{\theta}(Cx)^{-1/\alpha}).$$

The mean square angle is infinite now (due to the property of the density $g_2^{(\alpha)}$), but we may take

$$\delta\theta = (Cx)^{1/\alpha}$$

as the measure of the angular distribution width. This formula reveals a more rapid broadening of the angular distribution than in the normal case (15). We see here a superdiffusion regime considered in [1] (a more detailed consideration for N-dimensional space is performed in [14]).

When we have the opposite case, $\alpha = 2$ and $\beta < 1$, the mean square root of the scattering angle θ exists and is

$$\langle \theta^2 \rangle^{1/2} = \left(\frac{4C}{\beta\Gamma(\beta)}\right)^{1/2} x^{\beta/2}, \quad \beta < 1.$$

This is a subdiffusion [15].

In the general case the rate of broadening the angular distribution is determined by the exponent β/α:

$$\delta\theta = C^{1/\alpha} x^{\beta/\alpha}.$$

Thus the multiple scattering reveals the subdiffusion asymptotic behaviour if $\beta/\alpha < 1/2$ and superdiffusion one if $\beta/\alpha > 1/2$. Both the regimes are covered by term "anomalous diffusion" (see [8,16] and references therein).

However, not only the rate of broadening the angular distribution changes by passing from homogeneous medium to fractal one: the very shape of the density changes too. Whereas $\Psi_2^{(2,1)}(\vec{\theta})$ is simply the two-dimensional Gaussian distribution, the densities $\Psi_2^{(\alpha,1)}(\vec{\theta})$, $\alpha < 2$, are two-dimensional symmetrical stable densities with characteristic exponent α. They have a dome-shaped form near zero with the maximum value

$$\Psi_2^{(\alpha,1)}(0) = \frac{\Gamma(1+2/\alpha)}{4\pi},$$

48

long tails of a power type

$$\Psi_2^{(\alpha,1)}(\vec{\theta}) \sim \left[\frac{\Gamma(1+\alpha/2)}{2\pi}\right]^2 \sin(\alpha\pi/2)\theta^{-\alpha-2}, \quad \theta \to \infty$$

and no moments of the order $\geq \alpha$ (see [10]). On the contrary, the densities $\Psi_2^{(2,\beta)}(\vec{\theta})$, $\beta < 1$, diverge logarithmically at zero

$$\Psi_2^{(2,\beta)}(\vec{\theta}) \sim \frac{|\ln\theta|}{2\pi\Gamma(1-\beta)}, \quad \theta \to 0.$$

rapidly falling tails

$$\Psi_2^{(2,\beta)}(\vec{\theta}) \sim \frac{\beta^{(3\beta/2-1)/(2-\beta)}}{4\pi\sqrt{2-\beta}} \times$$

$$\times (\theta/2)^{-2(1-\beta)/(2-\beta)} \exp\left\{-(2-\beta)\beta^{\beta/(2-\beta)}(\theta/2)^{2/(2-\beta)}\right\}, \quad \theta \to \infty \qquad (16)$$

and finite moments of all positive orders

$$\int\limits_{\mathbb{R}^2} \theta^{2n}\Psi_2^{(2,\beta)}(\vec{\theta})d\vec{\theta} = 4^n\frac{[\Gamma(n+1)]^2}{\Gamma(\beta+1)}.$$

Note that (16) reduces to (14) when $\beta \to 1$.

In the case $\alpha = 1$, $\beta = 1/2$, the density $\Psi_2(\vec{\theta})$ is expressed in terms of the incomplete gamma function:

$$\Psi_2^{(1,1/2)}(\vec{\theta}) = (4\pi)^{-3/2}e^{\theta^2/4}\Gamma(-1/2,\theta^2/4).$$

Acknowledgments

References

1. S.V.Maleyev, *Phys. Rev. B*, **52**, 13163 (1995).
2. V.V. Uchaikin, D.A. Korobko, *Techn. Physics Letters*, **25**, 435 (1999).
3. V.V.Uchaikin, G.G.Gusarov, *J.Math.Phys.*, **38**, 2453 (1997).
4. V.Uchaikin, I.Gismjatov, G.Gusarov, V.Svetukhin, *Intern. J. Bif. and Chaos*, **8**, 977 (1998).
5. V.V.Uchaikin, G.G.Gusarov, D.A.Korobko, *Journ. Math. Sci.*, **92**, 3940 (1998).
6. E.W.Montroll, G.H.Weiss, *J.Math. Phys.*, **6**, 167 (1965).
7. V.V.Afanasiev, R.Z.Sagdeev, G.M.Zaslavsky, *Chaos* 1(2), 143 (1991).
8. V.V.Uchaikin, *Physica A* **255**, 65 (1998).
9. W.Feller, *Introduction to Probability Theory and Its Applications* (Vol. II, J.Wiley, New York, 1971).

10. S.G.Samko, A.A.Kilbas, O.I.Marichev, *Fractional Integrals and Derivatives – Theory and Applications* (Gordon and Breach, New York, 1993).
11. V.V.Uchaikin, V.M.Zolotarev, *Chance and Stability. Stable Distributions and their Applications* (VSP, Utrecht, The Netherlands, 1999).
12. V.M.Zolotarev, *One-dimensional Stable Distributions* (Amer. Math. Soc., Providence, R.I., 1986).
13. A.M. Mathai, R.K.Saxena, *The H-function with Applications in Statistics and Other Disciplines* (Wiley, New Delhi, 1978).
14. V.M.Zolotarev, V.V.Uchaikin, V.V.Saenko, *Journ. of Exper. and Theor. Physics*, **88**, 780 (1999).
15. V.V.Uchaikin, *Journ. of Exper. and Theor. Physics*, **88**, 1155 (1999).
16. J.-P.Bouchaud and A.Georges, *Phys. Rep.*, **195**, 127 (1990).

SYMMETRIC FRACTALS GENERATED BY CELLULAR AUTOMATA

A. BARBÉ AND F. VON HAESELER

Department of Electrical Engineering, KU Leuven,
Kardinaal Mercierlaan 94, 3001 Heverlee, Belgium
E-mail: {Andre.Barbe,fvanhaes}@esat.kuleuven.ac.be

We demonstrate how to construct fractals which are generated by a combination of a cellular automaton and a substitution. Moreover, if the substitution and the cellular automaton exhibit certain symmetry features, the fractal will inherit these symmetries.
Keywords: fractals, cellular automata, substitutions, symmetry, invariants

1 Introduction

It is a well known fact that certain cellular automata can generate fractals[7,8,9,14,15]. These fractals can be described in terms of hierarchical iterated function systems[12], graph directed constructions[11] or as a mixed self-similar set[2].

In this note we consider cellular automata which have certain symmetries. After the introduction of some basic concepts and formalism about substitutions, fractals and cellular automata in Sections 2 and 3, we shall show in Section 4 how to construct fractals generated by a cellular automaton such that the fractal inherits the symmetries of the cellular automaton.

2 Substitutions and fractals

In this section we introduce the concept of a substitution which is needed for our applications. We discuss how substitutions generate a compact subset which exhibits a hierarchical self-similar structure, and thus is usually a fractal set.

We will consider sequences $(f_j)_{j\in\mathbb{Z}}$, where f_j belongs to a finite commutative ring $\mathcal{R}$ with 0 and 1. It will be useful to represent this sequence as a formal Laurent series with coefficients in $\mathcal{R}$, as $\underline{f}$ or as

$$\underline{f} = f(X) = \sum_{j\in\mathbb{Z}} f_j X^j.$$

The set of all such Laurent series will be denoted by $\mathcal{R}(X)$. The support $\text{supp}(\underline{f})$ of $\underline{f} \in \mathcal{R}(X)$ is the set $\{j \mid f_j \neq 0\}$. $\underline{f}$ is called a Laurent polynomial if $\text{supp}(\underline{f})$ is finite. The set of all Laurent polynomials is denoted by $\mathcal{R}_c(X)$. If $\text{supp}(\underline{f}) \subset \mathbb{N}$ and is finite, then $\underline{f}$ is called a polynomial. The set of polynomials with coefficients in $\mathcal{R}$ is denoted by $\mathcal{R}[X]$.

The set of all maps $a : \mathcal{R} \to \mathcal{R}$ equipped with the addition and composition is denoted by $\text{Abb}(\mathcal{R})$. With 0 we denote the map $r \mapsto 0$ and with 1 we denote the identity map $r \mapsto r$. The subset of all maps $a : \mathcal{R} \to \mathcal{R}$ with $a(0) = 0$ is denoted as $\text{Abb}_0(\mathcal{R})$. If $a \in \text{Abb}(\mathcal{R})$, then a induces a map, also denoted by a, from $\mathcal{R}(X)$ to $\mathcal{R}(X)$ which is defined as $a(\underline{f})(X) = \sum_{j\in\mathbb{Z}} a(f_j)X^j$.

For $k \in \mathbb{N}$ the set $\{0,\ldots,k-1\}$ is denoted by $[k]$.

A k-substitution transforms a sequence $(f_j)_{j\in\mathbb{Z}}$ into a new sequence by replacing each element f_j by a string of k elements $\xi_0(f_j)\xi_1(f_j)\ldots\xi_{k-1}(f_j)$, where $\xi_j \in$ Abb($\mathcal{R}$). Here follows the formal

Definition 2.1 *A k-substitution ξ is a k-tuple $\xi = (\xi_l)_{l\in[k]} \in (\mathrm{Abb}(\mathcal{R}))^k$ which defines a map ξ on $\mathcal{R}(X)$ given by*

$$\xi(\underline{f}) = \sum_{l=0}^{k-1} X^l \xi_l(\underline{f})(X^k),$$

where $\xi_j(\underline{f})(X) = \sum_{j\in\mathbb{Z}} \xi_l(f_j)X^j$. If $\xi \in (\mathrm{Abb}_0(S))^k$, i.e., when 0 is replaced by a string of k 0's, then ξ is called a regular k-substitution.

Remark **1.** A k-substitution $\xi = (\xi_l)_{l\in[k]}$ is also defined by a polynomial $P_\xi(X) = \sum \xi_j X^j \in \mathrm{Abb}(\mathcal{R})[X]$ of degree less than k and we write $\xi(\underline{f}) = P_\xi(X) \circ f(X^k)$, where the product is the product induced by the composition of maps and the coefficients f_j of $f(X)$ are considered as constant maps $r \mapsto f_j$. The polynomial P_ξ is called the *substitution polynomial.*
2. If $Q(X) \in \mathcal{R}[X]$ is a polynomial of degree less than or equal to $k - 1$, then $\xi_Q(\underline{f}) = Q(X)f(X^k)$ defines a k-substitution with $\xi_j(r) = q_j r$.
3. If ξ_1 and ξ_2 are k-substitutions, respectively, and if $P_{\xi_1}(X)$ and $P_{\xi_2}(X)$ denote the respective substitution polynomials, then $\xi = \xi_1 \circ \xi_2$ is a k^2 substitution with substitution polynomial $P_\xi(X) = P_{\xi_1}(X) \circ P_{\xi_2}(X^k)$.

We also need to consider two-dimensional sequences $(g_{i,j})_{i,j\in\mathbb{Z}}$ with $g_{i,j} \in \mathcal{R}$. The corresponding formal Laurent series representation is

$$\underline{g} = f(X,Y) = \sum_{i,j\in\mathbb{Z}} g_{i,j}X^iY^j,$$

$\mathcal{R}(X,Y)$ is the set of all such two-dimensional Laurent series (sequences). Like for the one-dimensional case, we say that $g(X,Y)$ is a Laurent polynomial if $g_{i,j} = 0$ almost everywhere. Like for the one-dimensional case, two-dimensional sequences can be transformed by a two-dimensional $(k \times k)$-substitution, which replaces each symbol $g_{i,j}$ by a $(k \times k)$-array of symbols in $\mathcal{R}$. Formally:

Definition 2.2 *Let $\Xi = (\xi_{l,m})_{l,m\in[k]} \in (\mathrm{Abb}(\mathcal{R}))^{k\times k}$. The map $\Xi : \mathcal{R}(X,Y) \to \mathcal{R}(X,Y)$ defined as*

$$\Xi(\underline{g}) = \sum_{l,m=0}^{k-1} X^lY^m \xi_{l,m}(\underline{g})(X^k,Y^k)$$

is called a $(k \times k)$-substitution. It is called regular *if $\Xi \in (\mathrm{Abb}_0(\mathcal{R}))^{k\times k}$.*

Let us agree here, that in two-dimensional sequences and arrays, and in the corresponding graphical representations, the first index, which is possibly associated to the symbol X in formal Laurent series, increases along a horizontal axis which is oriented from the left to right; while the second index, possibly associated to the symbol Y, increases along a vertical axis which is oriented from top to bottom.

Remark **1.** If Ξ is a regular $(k \times k)$-substitution, then $\Xi(\mathcal{R}_c(X)) \subseteq \mathcal{R}_c(X)$. In other words, the set of Laurent polynomials is invariant under regular k- or $(k \times k)$-substitutions.

2. If Ξ_1 and Ξ_2 are $(k \times k)$-substitutions, respectively, and P_{Ξ_1}, P_{Ξ_2} the respective substitution polynomials, then $\Xi_1 \circ \Xi_2$ is a $(k^2 \times k^2)$-substitution with substitution polynomial $P_{\Xi_1 \circ \Xi_2}(X, Y) = P_{\Xi_1}(X, Y) \circ P_{\Xi_2}(X^k, Y^k)$.

In order to draw the connection between fractals and substitutions we define a graphical representation of elements in $\mathcal{R}_c(X)$ or in $\mathcal{R}_c(X, Y)$. We only need to introduce the graphical representation for elements in $\mathcal{R}_c(X, Y)$, since any Laurent series in $\mathcal{R}(X)$ has a natural counterpart in $\mathcal{R}(X, Y)$.

Let $(\mathcal{H}(\mathbb{R}^2), d_H)$ denote the set of non-empty compact subsets of $\mathbb{R}^2$, where d_H is the Hausdorff distance induced by the Euclidian metric (or any other equivalent metric) on $\mathbb{R}^2$. Since $\mathbb{R}^2$ is a complete metric space, so is $(\mathcal{H}, d_H)$[5,6].

Definition 2.3 *Let* $I(i, j) = [i, i + 1] \times [j, j + 1]$. *Then the map* $\mathbf{G} : \mathcal{R}_c(X, Y) \rightarrow \mathcal{H}(\mathbb{R}^2) \cup \{\emptyset\}$ *defined as*

$$\mathbf{G}(\underline{g}) = \{I(i, j) \mid g_{i,j} \neq 0\}$$

is called graphical representation.

$I(i, j)$ in the definition above can be considered as a "highlighted pixel" at location (i, j) on a graphical display which coincides with the $\mathbb{R}^2$-plane (as agreed: y-axis positively oriented downwards).

Remark   1. $\mathbf{G}(\underline{g}) = \emptyset$ if and only if $\underline{g} = 0$.
2. Any map $\gamma : \mathcal{R} \setminus \{0\} \rightarrow \mathcal{H}(\mathbb{R}^2)$ induces another graphical representation $\mathbf{G}_\gamma$ by defining $\mathbf{G}_\gamma(\underline{g}) = \cup \{(i, j) + \gamma(g_{i,j}) \mid g_{i,j} \neq 0\}$. This corresponds to " non-square pixels", where the pixel-shape may depend on $g_{i,j}$.
3. The concept of graphical representation is quite general: when $\mathcal{S}$ is a ring with 0 and 1, then so is $\mathcal{R} = \mathcal{S}^D$, with $0_{\mathcal{R}} = (0, 0, 0, \ldots, 0)$ and $1_{\mathcal{R}} = (1, 1, 1, \ldots, 1)$ (both $0_{\mathcal{R}}$ and $1_{\mathcal{R}}$ have D components). So it is possible to speak about the graphical representation of a two-dimensional sequence with values in $\mathcal{R} = \mathcal{S}^D$, where $g_{i,j}$ is just an element of $\mathcal{S}^D$.

The connection between fractals and substitutions is provided by the following theorem.

Theorem 2.4 [8] *Let* $\Xi : \mathcal{R}(X, Y) \rightarrow \mathcal{R}(X, Y)$ *be a regular* $(k \times k)$-*substitution such that for all* $r \neq 0$, *there are* i, j *such that* $\xi_{i,j}(r) \neq 0$. *If* $\underline{g} \in \mathcal{R}_c(X, Y)$ *and* $\underline{g} \neq 0$, *then the sequence*

$$\left(\frac{1}{k^n} \mathbf{G}(\Xi^n(\underline{g})) \right)_{n \in \mathbb{N}}$$

is a Cauchy sequence in $(\mathcal{H}, d_H)$.
The proof relies on two facts. Firstly, Ξ being regular maps Laurent polynomials on Laurent polynomials. Thus the above sequence is indeed in $\mathcal{H}(\mathbb{R}^2)$. Secondly, again due to the regularity of Ξ one has $\frac{1}{k} \mathbf{G}(\Xi(\underline{g})) \subset \mathbf{G}(\underline{g})$ for all Laurent polynomials different from 0. These two observations conclude the proof.

Remark   1. The limit of the above Cauchy sequence is denoted by $\mathcal{A}(\underline{g})$.
2. If $\mathbf{G}_\gamma$ is another graphical representation, then the above theorem remains true. Moreover, the limit set does not depend on the graphical representation.

3. For $\underline{g} \in \mathcal{R}_c(X,Y)$, $\underline{g} \not\equiv 0$, one has

$$\mathcal{A}(\underline{g}) = \bigcup_{\{(i,j) \mid g_{i,j} \neq 0\}} \mathcal{A}(g_{i,j}) + (i,j),$$

where $\mathcal{A}(g_{i,j}) + (i,j)$ is the translated limit set of the graphical representation of $g(X,Y) = g_{i,j}$.

Example 1. Let $\mathcal{R} = \mathbb{Z}_2$ and let the 3-substitution be given by the substitution polynomial $P_\xi(X) = \text{id} + \text{id}\, X^2$, where id denotes the 1 in Abb$(\mathcal{R})$. The limit set $\mathcal{A}(1)$, i.e. the limit of the properly rescaled graphical representation of the sequence $\cdots 0001000 \cdots$ under the substitution $0 \mapsto 000$, $1 \mapsto 101$ is the triadic Cantor set.

2. Let $\mathcal{R} = \mathbb{Z}_2$ and define the (2×2)-substitution by the substitution polynomial $P_\Xi(X,Y) = \text{id} + \text{id}\, Y + \text{id}\, XY$. The limit set $\mathcal{A}(1)$, i.e. the limit of the properly rescaled graphical representation of the sequence $g(X,Y) = 1$ (a single value 1 at $(i,j) = (0,0)$) under the substitution $0 \mapsto \begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}$, $1 \mapsto \begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}$, is the Sierpinski triangle.

3   Cellular automata

Substitutions transform a sequence into another sequence. So do other mechanisms known as cellular automata. We review briefly some essentials which we need further.

Let $\mathcal{S}$ denote a commutative ring with 0 and 1. According to Hedlund[10], a cellular automaton with states in the ring $\mathcal{S}$ (actually it is only necessary that $\mathcal{S}$ is a finite set) is a continuous map $A : \mathcal{S}(X) \to \mathcal{S}(X)$ which commutes with the shift map. This implies that every cellular automaton A is given by a local rule $\phi : \mathcal{S}^{N_1+N_2} \to \mathcal{S}$, where $N_1, N_2 \in \mathbb{N}$ and A is defined as

$$A(\underline{f}) = \sum_{j \in \mathbb{Z}} \phi(f_{j-N_1}, f_{j-N_1+1}, \ldots, f_{j+N_2-1}, f_{j+N_2}) X^j.$$

This means that the j-th element of $A(\underline{f})$ is obtained from $\underline{f}$ as a function ϕ of the $\underline{f}$-elements in a fixed finite neighbourhood of j. For a cellular automaton $A : \mathcal{S}(\overline{X}) \to \mathcal{S}(X)$ and an initial configuration $\underline{f} \in \mathcal{S}(X)$ we define the orbit $O_A(\underline{f})$ of $\underline{f}$ (w.r.t. A) as the two-dimensional sequence defined on $\mathbb{Z} \times \mathbb{N}$, where row $t \in \overline{\mathbb{N}}$ displays the t-th iterate of A on $\underline{f}$. Formally, and in Laurent series notation:

$$O_A(\underline{f})(X,Y) = \sum_{t=0}^{\infty} A^t(f(X)) Y^t.$$

In this note we are dealing with cellular automata $A = A_\phi$ with local rule $\phi : \mathcal{S}^D \to \mathcal{S}$ such that

$$A_\phi(\underline{f})(X) = \sum_{j \in \mathbb{Z}} \phi(f_{j-D+1}, \ldots, f_j) X^j.$$

Of particular interest are linear cellular automata[1], i.e., automata with a $\mathcal{S}$-linear local rule $\phi : \mathcal{S}^D \to \mathcal{S}$. In this case the cellular automaton is also described by the multiplication with a "local-rule"-polynomial $R(X)$, i.e., $A_\phi(\underline{f}) = R(X)f(X)$.

We are particularly interested in cellular automata with the additional property of being k-Fermat[9]. They provide a framework to explain the existence of fractals which are generated by cellular automata.

For $k \in \mathbb{N} \setminus \{0\}$ we define the *k-th power* $\pi_k : \mathcal{S}(X) \to \mathcal{S}(X)$ as $\pi_k(\underline{f})(X) = \sum f_j X^{jk} = f(X^k)$, i.e., π_k "inflates" $\underline{f}$ by inserting $(k-1)$ 0's between the elements of the sequence $\underline{f}$ (we could call this "k-inflation").

We now recall the concept of a cellular automaton having the k-Fermat property. This implies that row k of the orbit of a "k-inflated" initial sequence equals the "k-inflation" of row 1 of the orbit of the "non-inflated" initial sequence, no matter what the initial sequence might be. Formally:

Definition 3.1 [1,9] *A cellular automaton $A : \mathcal{S}(X) \to \mathcal{S}(X)$ is called k-Fermat if $(A^k \circ \pi_k)(\underline{f}) = (\pi_k \circ A)(\underline{f})$ holds for all $\underline{f} \in \mathcal{S}(X)$.*

Remark **1.** If A is a k-Fermat cellular automaton, then the sequence 0 is a fixed point of A. Therefore $A(\mathcal{S}_c(X)) \subseteq \mathcal{S}_c(X)$ for a k-Fermat cellular automaton.
2. If A is a linear cellular automaton defined by the Laurent polynomial $R(X)$, then A is k-Fermat if and only if $R(X)^k = R(X^k)$.

Example **1.** Let $\mathcal{S} = \mathbb{F}_{p^\alpha}$, the field with p^α elements and of characteristic p (p being a prime number). Then any linear cellular automaton is p^α-Fermat.
2. Let $\mathcal{S} = \mathbb{Z}_{p^\alpha}$, p a prime number. A linear cellular automaton is p-Fermat if its local rule is a Laurent polynomial $R(X)$ of the form $R(X) = Q(X)^{p^{\alpha-1}}$, where $Q(X)$ is a Laurent polynomial.
3. Let $\mathcal{S} = \mathbb{Z}_6$, then $R(X) = 3 + 4X$ defines a 2-Fermat linear cellular automaton and $R(X) = 2 + 4X$ defines a 3-Fermat linear cellular automaton.
4. Let $\mathcal{S} = \mathbb{Z}_5$. Then the local rule $\phi : \mathbb{Z}_5^2 \to \mathbb{Z}_5$ defined by the table with entries $\phi(x, y)$

$x \backslash y$	0 1 2 3 4
0	0 1 2 3 4
1	1 0 4 2 3
2	2 3 0 4 1
3	3 4 1 0 2
4	4 2 3 1 0

induces a (nonlinear) 2-Fermat cellular automaton. In fact, it can be shown that there are 5^{12} different local rules $\phi : \mathbb{Z}_5^2 \to \mathbb{Z}_5$ which define a 2-Fermat cellular automaton on $\mathbb{Z}_5(X)$. Which demonstrates that not all k-fermat cellular automata have to be linear cellular automata.

We now present a generic example which illustrates a general property of k-Fermat cellular automata. Consider the cellular automaton with $\mathcal{S} = \mathbb{Z}_3$ and local rule $R(X) = 1 + 2X + X^2 \in \mathbb{Z}_3[X]$. It is 3-Fermat. Part of the orbit $O_A(1)$ is shown below:

```
row 00     · · · 0001000000000000000000000000000000 · · ·
row 01     · · · 0001210000000000000000000000000000 · · ·
row 02     · · · 0001101100000000000000000000000000 · · ·
row 03     · · · 0001002001000000000000000000000000 · · ·
row 04     · · · 0001212121210000000000000000000000 · · ·
row 05     · · · 0001100000001100000000000000000000 · · ·
row 06     · · · 0001001000001001000000000000000000 · · ·
row 07     · · · 0001211210001211210000000000000000 · · ·
row 08     · · · 0001102201101102201100000000000000 · · ·
row 09     · · · 0001000000002000000001000000000000 · · ·
row 10     · · · 0001210000002420000001210000000000 · · ·
row 11     · · · 0001101100002202200001101100000000 · · ·
row 12     · · · 0001002001002001002001002001000000 · · ·
row 13     · · · 0001212121212121212121212121210000 · · ·
```

Row $3n + l$, $l \in \{0, 1, 2\}$ in this orbit, i.e. the sequence $A^{3n+l}(1)$, can be obtained from row $3n$ as $A^l(A^{3n}(1))$. By the 3-Fermat-property, row $3n$ is a "3-inflated" version of row n (say $\cdots zabcd \cdots$), as given by the top sequence in the following scheme:

$$
\begin{array}{llllllll|lll|llllll}
\text{row } 3n & \cdots 0 & 0 & z & 0 & 0 & a & 0 & \mathbf{0} & \mathbf{b} & \mathbf{0} & \mathbf{0} & c & 0 & 0 & d & 0 & 0 \cdots \\
\text{row } 3n+1 \cdots & & & 2z & z & a & 2a & \mathbf{a} & \mathbf{b} & 2\mathbf{b} & \mathbf{b} & c & 2c & c & d & 2d & d \cdots \\
\text{row } 3n+2 \cdots & & & & a+z & a+z & \mathbf{0} & \mathbf{a+b} & \mathbf{a+b} & \mathbf{0} & b+c & b+c & 0 & c+d & c+d & 0 \cdots
\end{array}
$$
$$(1)$$

Also represented in this scheme are rows $3n+1$ and $3n+2$, obtained by application of the rule $R(X)$. Observe from this scheme that knowledge of two successive elements in row n, say (a, b), is sufficient to determine all highlighted values. Now define the substitution Ψ which replaces the pair (a, b) by a (3×3)- array of pairs, as follows:

$$
(a, b) \mapsto \begin{bmatrix}
(0, b) & (b, 0) & (0, 0) \\
(a, b) & (b, 2b) & (2b, b) \\
(0, a+b) & (a+b, a+b) & (a+b, 0)
\end{bmatrix}.
$$

The last components of the pairs in this array are given by the elements of the 3×3- array indicated in the above three sequences (1); the first components are the elements that immediately precede them. It is now clear that, if we replace each element (say b) in row n by the corresponding last components in row l ($l = 0, 1, 2$) of the substitution array, we produce row $3n + l$. In that way, the orbit above corresponds to the last elements of the pairs in the arrays obtained by iterating the substitution starting from the sequence of pairs $\cdots (0, 0)\ (0, 1)\ (1, 0)\ (0, 0) \cdots$ (i.e., row 0 rewritten as pairs in which the last components correspond to the elements of row 0, and the first components to the corresponding preceeding elements). We show the first step of the iteration (elements of the actual orbit are underlined):

$$\cdots (0,\underline{0})(0,\underline{1})(1,\underline{0})(0,\underline{0}) \cdots$$

$$\downarrow$$

$$\cdots \left| \begin{array}{c} (0,\underline{0})(0,\underline{0})(0,\underline{0}) \\ (0,\underline{0})(0,\underline{0})(0,\underline{0}) \\ (0,\underline{0})(0,\underline{0})(0,\underline{0}) \end{array} \right| \begin{array}{c} (0,\underline{1})(1,\underline{0})(0,\underline{0}) \\ (0,\underline{1})(1,\underline{2})(2,\underline{1}) \\ (0,\underline{1})(1,\underline{1})(1,\underline{0}) \end{array} \left| \begin{array}{c} (0,\underline{0})(0,\underline{0})(0,\underline{0}) \\ (1,\underline{0})(0,\underline{0})(0,\underline{0}) \\ (0,\underline{1})(1,\underline{1})(1,\underline{0}) \end{array} \right| \begin{array}{c} (0,\underline{0})(0,\underline{0})(0,\underline{0}) \\ (0,\underline{0})(0,\underline{0})(0,\underline{0}) \\ (0,\underline{0})(0,\underline{0})(0,\underline{0}) \end{array} \right| \cdots$$

In the next step, each pair is again subjected to the substitution Ψ. In general, the pairs will be D-tuples, and the substitution will replace D-tuples by $(k \times k)$-arrays of D-tuples (for a k-Fermat cellular automaton). We now express this property as a formal theorem.

Let D be a natural number, $D \geq 1$. We define an embedding $\iota = \iota_D$ of $\mathcal{S}(X)$ into $\mathcal{S}^D(X)$ as follows ($\mathcal{S}^D(X)$ can be considered as the set of sequences with D-tuples as values). For $f(X) = \sum_{j\in\mathbb{Z}} f_j X^j \in \mathcal{S}(X)$ we define $\iota(\underline{f}) \in \mathcal{S}^D(X)$ as

$$\iota(\underline{f})(X) = \sum_{j\in\mathbb{Z}} (f_{j-D+1}, \ldots, f_j) X^j.$$

We also define a projection $\mathbf{p} = \mathbf{p}_D : \mathcal{S}^D(X) \to \mathcal{S}(X)$ such that $\mathbf{p} \circ \iota(\underline{f}) = \underline{f}$ by setting $\mathbf{p}\left(\sum_{j\in\mathbb{Z}}(s_{j,-D+1},\ s_{j,-D+2}, \ldots, s_{j,0}) X^j\right) = \sum_{j\in\mathbb{Z}} s_{j,0} X^j$.

With these notions the work of several authors[7,8,9,13,14] on fractals generated by cellular automata can be summarized as

Theorem 3.2 *If* $A : \mathcal{S}(X) \to \mathcal{S}(X)$ *is a k-Fermat cellular automaton, then there exists a $D \in \mathbb{N}$ and a regular $(k \times k)$-substitution $\Psi = (\psi_{i,j})_{i,j\in[k]} : \mathcal{S}^D(X,Y) \to \mathcal{S}^D(X,Y)$ such that for all $n \in \mathbb{N}$ and $l \in [k]$ we have*

$$A^{nk+l}(1) = \mathbf{p}\left(\sigma_l(\iota(A^n(1)))\right),$$

where σ_l is the k-substitution $(\psi_{i,l})_{i\in[k]}$.

The interpretation in terms of fractals generated by cellular automata is clear. If A is a k-Fermat cellular automaton, then the initial configuration 1 generates an orbit. If the orbit is visualized by a graphical representation, one observes a certain pattern. According to Theorem 3.2, the pattern can be described by a $(k \times k)$-substitution $\Psi = (\psi_{l,m})$ and by Theorem 2.4 the pattern is represented by a compact subset. Moreover, for linear k-Fermat cellular automata it is known that the limit set does not dependent on the initial configuration $\underline{f} \in \mathcal{S}_c(X)$.

4 Cellular automata, symmetries and invariants

It is worthwhile to study cellular automata with a local rule $\phi : \mathcal{S}^2 \to \mathcal{S}$ which exhibits certain symmetries [3,4].

Definition 4.1 *The map $\phi : \mathcal{S}^2 \to \mathcal{S}$ is called* rotationally symmetric *if $\phi(\phi(x,y),x) = y$ holds for all $x, y \in \mathcal{S}$. If ϕ is rotationally symmetric and commutative, i.e., $\phi(x,y) = \phi(y,x)$, then ϕ is called* totally symmetric.

58

The following geometric interpretation clarifies the definition. We represent the relation $z = \phi(x, y)$ as a triangular array $^x z^y$, and call it an elementary ϕ-configuration. In geometric terms the rotational symmetry of ϕ can be phrased as: Any rotation (by $\frac{2\pi}{3}$) of an elementary ϕ-configuration, changing $^x z^y$ into $^z y^x$ or $^y x^z$, is an elementary ϕ-configuration.

The commutativity of ϕ implies the invariance of elementary ϕ-configurations under vertical reflection. Combined with the rotational symmetry of ϕ we can state that ϕ is totally symmetric if and only if the set of elementary ϕ-configurations is invariant under the symmetry group of the equilateral triangle.

Example **1.** The local rule $\phi : \mathbb{Z}_2^2 \to \mathbb{Z}_2$ defined as $\phi(x, y) = x + y$ is totally symmetric.

2. Let $\mathbb{F}_4 = \{0, 1, \zeta, 1 + \zeta\}$ denote the field with 4 elements. The local rule $\phi : \mathbb{F}_4^2 \to \mathbb{F}_4$ defined as $\phi(x, y) = \zeta x + (1 + \zeta)y$ is rotationally symmetric but not totally symmetric.

3. There exist rotationally symmetric local rules which are not related to linear rules.[4]

Definition 4.2 *A ϕ-configuration of size N is a top down equilateral triangular array with N elements in the top row and with values in S such that any subtriangle $^{xy}_{\ z}$ is an elementary ϕ-triangle.*

A ϕ-configuration of size N is rotationally symmetric if the rotated ϕ-configuration is equal to the ϕ-configuration.

A ϕ-configuration of size N is totally symmetric if it remains unchanged under the symmetry group of the equilateral triangle.

Figure 1 shows examples of ϕ-configurations.
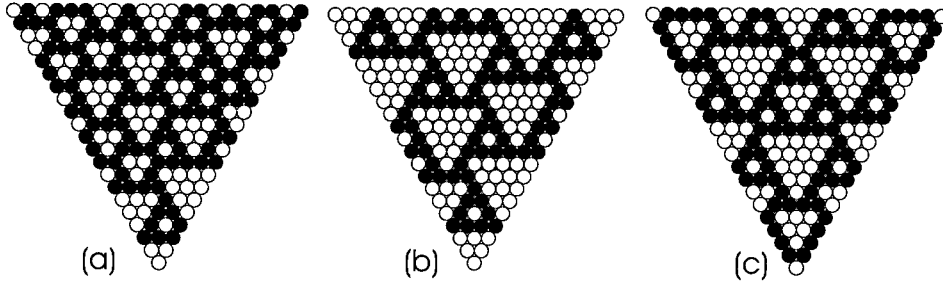
Figure 1. Examples of ϕ-configurations of size 21 for $\phi(x, y) = x + y$ defined in $\mathbb{Z}_2$. (a): a nonsymmetric one, (b): a rotationally symmetric one, (c): a totally symmetric one. (A black cell represents the value 1, a white cell the value 0).

Theorem 4.3 [4] *If $\phi : S^2 \to S$ is rotationally symmetric and of the form $\phi(x, y) = rx + sy$, then there exist for any $N \in \mathbb{N}$ a rotationally symmetric ϕ-configuration of size N.*

If $\phi : S^2 \to S$ is totally symmetric and of the form $\phi(x, y) = rx + sy$, then there exists for any $n \in \mathbb{N}$ a totally symmetric ϕ-configuration of size N.

As a next step we introduce invariants of cellular automata. The motivating example for invariants is given by the k-Fermat cellular automata. Note that the map $\pi_k : \mathcal{S}(X) \to \mathcal{S}(X)$ is a special kind of k-substitution. In fact, the substitution polynomial associated with π_k is $P_{\pi_k}(X) = \mathrm{id}$ and $A^k \circ \pi_k = \pi_k \circ A$. It is therefore natural to define

Definition 4.4 *Let $A : \mathcal{S} \to \mathcal{S}$ be a cellular automaton. A k-substitution ξ is called k-invariant of A if $A^k \circ \xi = \xi \circ A$. The set of all k-invariants of the cellular automaton A is denoted by $\mathrm{Inv}_k(A)$. The substitution ξ is called a regular k-invariant of A if ξ is a k-invariant and regular. The set of regular k-invariants is denoted by $\mathrm{Inv}_k^r(A)$.*

Remark *If $k = 1$, then the 1-invariants of a cellular automaton A with local rule $\phi : \mathcal{S}^D \to \mathcal{S}$ are given by $\mathrm{Inv}_1(A) = \{\zeta : \mathcal{S} \to \mathcal{S} \mid \phi(\zeta(r_1), \ldots, \zeta(r_D)) = \zeta(\phi(r_1, \ldots, r_M)), r_1, \ldots, r_D \in \mathcal{S}\}$.*

For linear k-Fermat cellular automata the coefficients of the substitution polynomial of a k-invariant ξ can be characterized via the local rule.

Theorem 4.5 *If A is a linear k-Fermat cellular automaton, then $\mathrm{Inv}_k(A) = \prod_{l=1}^{k} \mathrm{Inv}_1(A)$.*

Proof: Let $\underline{f} \in \mathcal{S}(X)$ and let $R(X) = \sum_{j=0}^{D} r_j X^j$ be the local rule of A. Since ξ is a k-invariant, we have that $R(X)^k \xi(\underline{f}) = \xi(R(X)\underline{f}(X))$. Since R is k-Fermat, the left side becomes $R(X^k)\xi(\underline{f}) = R(X^k) \sum_{j=0}^{(k-1)} X^j(\underline{f})(X^k)$ (using the definition of ξ). Thus the coefficient of $X^{kl+j}, j \in [k]$, in $R(X^k)\xi(\underline{f})$ can be computed as $r_0 \xi_j(f_l) + r_1 \xi_j(f_{l-1}) + \cdots + r_D \xi_j(f_{l-D})$. On the other hand, the same coefficient in $\xi(R(X)f(X))$ turns out to be $\xi_j(r_0 f_l + \cdots + r_D f_{l-D})$, i.e. $\xi_j \circ A = A \circ \xi_j$. Therefore, ξ is a k-invariant if and only if each ξ_j is a 1-invariant of A. $\bullet$

Example 1. For $\mathcal{S} = \mathbb{Z}_2$ and $R(X) = 1 + X$, we have $\mathrm{Inv}_1(A) = \{s \mapsto s, s \mapsto 0\}$ $(s \in \mathcal{S})$ and $\mathrm{Inv}_1(A) = \mathrm{Inv}_1^r(A)$. The automaton specified by this rule is 2^n-Fermat (for any $n \in \mathbb{N}$). Thus $\mathrm{Inv}_2^r(A)$ contains 2^2 regular 2-invariant substitutions : $0 \mapsto 00, 1 \mapsto 00; 0 \mapsto 00, 1 \mapsto 10; 0 \mapsto 00, 1 \mapsto 01$ and $0 \mapsto 00, 1 \mapsto 11$. $\mathrm{Inv}_4^r(A)$ contains 2^4 regular 4-invariants, among which $0 \mapsto 0000, 1 \mapsto 1011$.

2. For $\mathcal{S} = \mathbb{Z}_3$ and $R(X) = 2 + 2X$, we have $\mathrm{Inv}_1(A) = \{s \mapsto \alpha s + \beta \mid \alpha, \beta \in \mathbb{Z}_3\}$ and $\mathrm{Inv}_1^r(A) = \{s \mapsto \alpha s \mid \alpha \in \mathbb{Z}_3\}$. $\mathrm{Inv}_3^r(A)$ contains 3^3 regular k-invariants, among which $0 \mapsto 000, 1 \mapsto 201, 2 \mapsto 102$.

3. $R(X) = 1 + 3X \in \mathbb{Z}_6[X]$ defines a 2-Fermat automaton such that $|\mathrm{Inv}_1(A)| = 54$ and $|\mathrm{Inv}_1^r(A)| = 18$. Thus there are 2^{18} regular 2-invariants.

In analogy with the substitution Ψ introduced before Theorem 3.2, we associate here with a linear k-Fermat automaton A, given by $R(X) = s + rX$, and a k-invariant $\xi \in \mathrm{Inv}_k^r(A)$, a $(k \times k)$-substitution $\Psi_\xi = \Psi(A, \xi) : \mathcal{S}^2(X, Y) \to \mathcal{S}^2(X, Y)$. Let $(a, b) \in \mathcal{S}^2$ and consider the polynomials $Q_m(X) = R(X)^m \xi(aX^{-1} + b) = \sum_{j \in \mathbb{Z}} q_{j,m} X^j$ for $m \in [k]$. For $m \in [k]$ the coefficients $q_{-1,m}, q_{0,m}, \ldots, q_{k-1,m}$ of $Q_m(X)$ are uniquely determined by a and b, thus these coefficients can be regarded as functions from $\mathcal{S}^2$ to $\mathcal{S}$. Therefore, the substitution $\Psi_\xi = (\psi_{l,m})_{l,m \in [k]}$, defined as

$$\psi_{l,m} : \mathcal{S}^2 \to \mathcal{S}^2$$
$$(a, b) \mapsto (q_{l-1,m}(a, b), q_{l,m}(a, b))$$

for $l, m \in [k]$, is well defined and called the *induced* substitution.

Remark 1. If ξ is a regular k-invariant, then the induced substitution Ψ_ξ is a regular substitution.

2. If the local rule of a linear cellular automaton is given by the polynomial $R(X) \in \mathcal{S}[X]$, of degree d, and if ξ is a regular k-invariant, then the induced substitution is defined by maps from $\mathcal{S}^{d+1}$ which are constructed in a similar manner as for the case $d = 1$ considered above. $(d + 1)$-tuples are then replaced by a $(k \times k)$-array of $(d + 1)$-tuples.

3. If A_ϕ is any cellular automaton such that

$$A_\phi\left(\sum f_j X^j\right) = \sum \phi(f_{j-d}, \dots, f_j) X^j$$

and with a k-invariant ξ, then the above construction yields an induced substitution as well.

Example For $\mathcal{S} = \mathbb{Z}_2$, the polynomial $R(X) = 1 + X$ defines a 2^2-Fermat linear cellular automaton A. By Example 1 following Theorem 4.5, we have that the substitution ξ defined by the polynomial $P_\xi(X) = \mathrm{id} + 0X + \mathrm{id}X^2 + \mathrm{id}X^3$ defines a 4-invariant of A. Consider a sequence $\cdots zabcd \cdots$, then substitution ξ transforms this into the top row shown below:

$$
\begin{array}{llll|lll|lll}
\cdots z\,0\,z\,z & a & 0\,a\,a & \mathbf{b} & \mathbf{0}\,\mathbf{b}\,\mathbf{b} & c & 0\,c\,c \cdots \\
\cdots \quad z\,z\,0\,z+a & a\,a\,0 & a+b\,\mathbf{b}\,\mathbf{b}\,\mathbf{0} & b+c\,c\,c\,0 \cdots \\
\cdots \quad\quad 0\,z\,z+a & z\,0\,a & a+b\,\mathbf{a}\,\mathbf{0}\,\mathbf{b} & b+c\,b\,0\,c \cdots \\
\cdots \quad\quad\quad z & a & a\,z\,a & \mathbf{b} & \mathbf{b}\,\mathbf{a}\,\mathbf{b} & c & c\,b\,c \cdots
\end{array}
$$

The other rows are part of the orbit generated with the given cellular automaton rule starting from the first row. The induced substitution Ψ_ξ can be read directly from these rows:

$$
(a, b) \mapsto
\begin{bmatrix}
(a, b) & (b, 0) & (0, b) & (b, b) \\
(0, a+b) & (a+b, b) & (b, b) & (b, 0) \\
(a, a+b) & (a+b, a) & (a, 0) & (0, b) \\
(a, b) & (b, b) & (b, a) & (a, b)
\end{bmatrix}.
\tag{2}
$$

The following theorem is a straightforward generalization of Theorem 3.2. It exploits the existence of k-invariants instead of the k-Fermat property.

Theorem 4.6 *Let A_ϕ be a cellular automaton and let ξ be a regular k-invariant and let $\Psi_\xi = (\psi_{i,j})_{i,j \in [k]}$ be the induced k-substitution. There exists a natural number D, and embedding $\iota = \iota_D$, such that for $\underline{f} \in \mathcal{S}(X)$ with $\xi(\underline{f}) = \underline{f}$ and all $n \in \mathbb{N}$, $l \in [k]$ we have*

$$A^{nk+l}(\underline{f}) = \mathbf{p}\left(\sigma_l(\iota(A^n(\underline{f})))\right),$$

where σ_l is the k-substitution $(\psi_{i,l})_{i \in [k]}$.

Proof. Since ξ is a k-invariant and $\underline{f} = \xi(\underline{f})$, we have $A^{kn}(\underline{f}) = A^{kn}(\xi(\underline{f})) = \xi(A^n(\underline{f}))$ for all $n \in \mathbb{N}$. That proves the assertion for $l = 0$. The assertion for $l \neq 0$ follows from the definition of the induced substitution. $\bullet$

We illustate this theorem by continuing the above example, with $\mathcal{S} = \mathbb{Z}_2$,

$R(X) = 1 + X$ and the 4-invariant which produces the following substitutions: $0 \mapsto 0000$, $1 \mapsto 1011$. Take $\underline{f}$ as the sequence which is the limit obtained from applying this substitution to $\cdots 0001011000 \cdots$ (the boldface element is at position 0). This yields the top row in

$$\begin{aligned}
\underline{f} \quad & \cdots 000101100001011101100000000000000010111011 \cdots \\
A(\underline{f}) \quad & \cdots 000111010001110011010000000000000011100110 \cdots
\end{aligned}$$

Observe that $\underline{f} = \xi(\underline{f})$. Theorem 4.6 states that, for example, row 6 $(= 1 \cdot 4 + 2)$ in the orbit $O_A(\underline{f})$ can be obtained from $A(\underline{f})$ displayed above, by imbedding this last sequence in a sequence of pairs (corresponding to $D = 2$), as given by

$$\cdots (0,\underline{1})(1,\underline{1})(1,\underline{1})(1,\underline{0})(0,\underline{1})(1,\underline{0})(0,\underline{0})(0,\underline{0})(0,\underline{1})(1,\underline{1})(1,\underline{1}) \cdots$$

(the last (underlined) components in these pairs are the elements of $A(\underline{f})$, the first components are the ones that immediately precede them). Then apply the substitutions displayed in row 2 of the induced substitution array Ψ_ξ given above (see(2): row 2 is actually the third row, as counting starts at zero). This gives

$$\cdots (0,\underline{1})(1,\underline{0})(0,\underline{0})(0,\underline{1})|(1,\underline{0})(0,\underline{1})(1,\underline{0})(1,\underline{1})|(1,\underline{0})(0,\underline{1})(1,\underline{0})(1,\underline{1})|(1,\underline{1})(1,\underline{1}) \cdots$$

The last components in these pairs form row 6 in $O_A(\underline{f})$. In a similar way, using row l of the Ψ_ξ-substitution array, where $l \in \{0,1,2,3\}$, starting from the embedding of row $A^n(\underline{f})$ in a sequence of pairs, would produce row $4 \cdot n + l$ in this orbit.

As a consequence of Theorem 2.4, we obtain

Corollary 4.7 *If ξ is a regular k-invariant of the cellular automaton A and if Ψ_ξ is the induced substitution, then the sequence $\left(\frac{1}{k^n} \mathbf{G}(\Psi_\xi^n(\iota(\underline{f}))) \right)_{n \in \mathbb{N}}$ is a Cauchy sequence for all $\underline{f} \in \mathcal{S}_c(X)$.*

As a final step, we combine the notion of symmetry of a cellular automaton with the existence of induced substitutions. We are interested in the following problem. Suppose $f_0, \ldots, f_{N-1}$ is the top row of a ϕ-configuration of size N and assume that the configuration is, e.g., rotationally symmetric. If ξ is a regular k-invariant of A_ϕ, then $\xi(\sum f_j X^j)$ gives anther top row of a ϕ-configuration of size kN; what are the symmetry properties of the larger configuration?

Definition 4.8 *Let A_ϕ be a cellular automaton with local rule $\phi : \mathcal{S}^2 \to \mathcal{S}$ and let the substitution $\xi = (\xi_l)_{l \in [k]}$ be a k-invariant of A. The k-invariant is* palindromic *if $\xi_i = \xi_{k-i-1}$ holds for all $i \in [k]$.*

The k-invariant ξ is called rotationally symmetric *if the induced substitution $\Psi_\xi = (\psi_{l,m})$ satisfies*

$$\mathbf{p}(\psi_{i,0}(a,b)) = \mathbf{p}(\psi_{k-1,i}(a,b))$$

for all $(a,b) \in \mathcal{S}^2$ and all $i \in [k]$.

If $\xi \in \mathrm{Inv}_k(A)$ is palindromic and rotationally symmetric, then ξ is called totally symmetric.

An example of a rotationally symmetric 4-invariant is given by the example preceeding Theorem 4.6: the last components of the rightmost column in the Ψ_ξ-array equals the last components of its top row.

Remark If $\xi = (\xi_l)$ is a k-invariant for A_ϕ and $\Psi_\xi = (\psi_{l,m})$ the induced substitution, then we have $\mathbf{p}(\psi_{i,0}(a,b)) = \xi_i(b)$ for $i \in [k]$ and there exist functions $\rho_i \in \mathrm{Abb}(\mathcal{S})$, $i \in [k]$, such that $\mathbf{p}(\psi_{k-1,i}(a,b)) = \rho_i(b)$. In particular, if ξ is rotationally symmetric, then we have $\rho_i = \xi_i$ for all $i \in [k]$.

Lemma 4.9 [4] *Let A be a cellular automaton with local rule $\phi : \mathcal{S}^2 \to \mathcal{S}$. If $f(X) \in \mathcal{S}[X]$ is of degree $\leq N$, then $f(X)$ defines a rotationally symmetric ϕ-configuration of size $(N+1)$ if and only if for the coefficients $O_{i,j}$ of $O_A(\underline{f})(X,Y)$ the equations $O_{i,0} = O_{N,i}$ hold for $i = 0, \ldots, N$.*

In the linear case a carefully performed substitution can preserve symmetries.

Lemma 4.10 *Let A_ϕ be a cellular automaton with rotationally symmetric local rule $\phi(x,y) = rx + sy$. If $f_0, \ldots, f_{N-1}$ is a top row of a rotationally symmetric ϕ-configuration of size N and if $\xi \in \mathrm{Inv}_k^{\mathrm{r}}(A_\phi)$ is rotationally symmetric, then $g_0, \ldots, g_{kN-1}$, where $g(X) = \sum_{j \in \mathbb{Z}} g_j X^j = \xi(\sum_{i=0}^{N-1} f_i X^i)$, is the top row of a rotationally symmetric ϕ-configuration of size kN.*

Proof: Let $f(X) = \sum_{j=0}^{N-1} f_j X^j$ and $\underline{Q} = O_A(\underline{f})(X,Y) = \sum O_{i,j} X^i Y^j$ its orbit. If Ψ_ξ denotes the induced $(k \times k)$-substitution, then $\Psi_\xi(\underline{Q}) = \sum \Theta_{i,j} X^j Y^j$ is the orbit of $\xi(\underline{f})$. It remains to prove that $\Theta_{i,0} = \Theta_{kN-1,i}$ holds for all $i = 0, \ldots, kN - 1$.

By Lemma 4.9, we have $O_{i,0} = O_{N-1,i}$ for $i = 0, \ldots, N-1$. Since ξ is rotationally symmetric, we have $\Theta_{kj+l,0} = \xi_l(O_{j,0})$ for $kj + l \in [Nk]$ and $l \in [k]$ and $\Theta_{kN-1,kj+l} = \xi_l(O_{N-1,j}) = \xi_l(O_{j,0})$, which proves the assertion. $\bullet$

Lemma 4.10 remains true if rotationally symmetric is replaced by totally symmetric.

We can now establish the existence of fractals with prescribed symmetries.

Theorem 4.11 *Let A_ϕ, with $\phi(x,y) = rx + sy$, define a cellular automaton with a rotationally (totally) symmetric linear local rule. Let $f(X) \in \mathcal{S}[X]$ of degree $\leq N$ define a rotationally (totally) symmetric ϕ-configuration of size $(N+1)$ and let $\xi \in \mathrm{Inv}_k^{\mathrm{r}}(A_\phi)$ be a regular rotationally (totally) symmetric k-invariant. If $\iota(O_A(\underline{f})) = \iota_2(O_A(\underline{f}))$ denotes the embedding of the orbit $O_A(\underline{f})$ which replaces $(O_A(\underline{f}))_{j,i}$ by $((O_A(\underline{f}))_{j-1,i}, (O_A(\underline{f}))_{j,i})$, and if Ψ_ξ denotes the induced substitution, then the sequence*

$$\left(\frac{1}{k^n (N+1)} \begin{pmatrix} 1 & -1/2 \\ 0 & \sqrt{3}/2 \end{pmatrix} \mathbf{G}(\Psi_\xi^n(\iota(O_A(\underline{f})))) \cap \nabla \right)_{n \in \mathbb{N}},$$

where ∇ is the equilateral triangle given by the points $(0,0)$, $(1,0)$ and $(1/2, \sqrt{3}/2)$ (ordinate axis positively oriented downwards), is a Cauchy sequence. The limit, denoted as $\mathcal{X} = \mathcal{X}(\underline{f}, \xi)$, is a compact set in ∇ and $\mathcal{X}$ is rotationally (totally) symmetric.

Proof: By Theorem 2.4, the limit $1/k^n \mathbf{G}(\Psi_\xi^n(\iota(\underline{f})))$ exists and is contained in $[0, N+1]^2$. The factor transforms the triangle $\{(x,y) \in [0, N+1]^2 \mid y \leq x\}$ into an equilateral triangle of size 1. Since the limit is independent of the particular graphical representation ("pixel-shape"), we can define $\gamma : \mathcal{S}^2 \setminus \{0\} \to \mathcal{H}(\mathbb{R}^2)$ by setting $\gamma(s_{-1}, s_0) = \cup\{(j,0) \mid s_j \neq 0; j = -1, 0\}$. Then the rescaled version of $G_\gamma(\Psi_\xi^n(\iota(\underline{f}))) \cap \nabla$ is rotationally (totally) symmetric for each $n \in \mathbb{N}$. Thus the limit is rotationally (totally) symmetric. $\bullet$
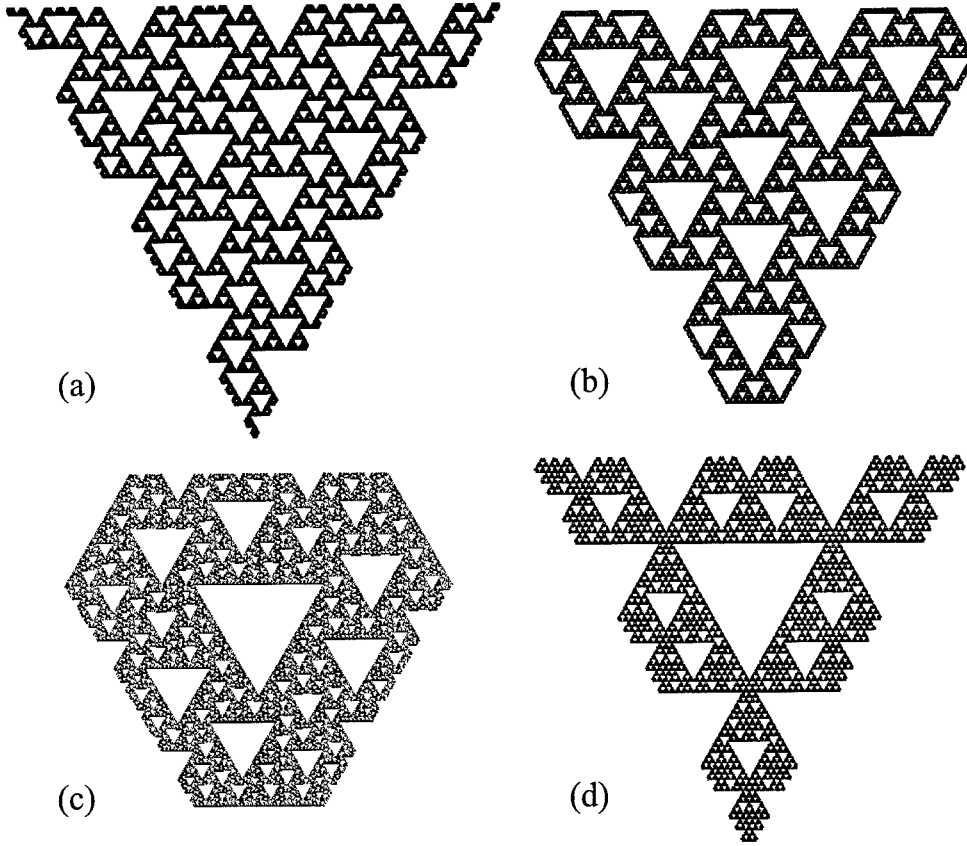
Figure 2. Symmetric fractals generated by cellular automata. (a) and (b): states in $\mathbb{Z}_2 = \{0,1\}$ with local rule $\phi(x,y) = x + y$ and, for (a): with top row generated from the initial top-row configuration 1101101101 and the 16-substitution $0 \mapsto$ 0000000000000000, $1 \mapsto$ 0110011000000110,: for (b): initial configuration 0110110110 and 8-substitution $0 \mapsto$ 00000000, $1 \mapsto$ 01111110. (c): states in $\mathbb{F}_4 = \{0,1,\zeta,1+\zeta\} = \{0,1,2,3\}$ with local rule $\phi(x,y) = \zeta x + (1 + \zeta)y$, initial state 0130 and 4-substitution $0 \mapsto$ 0000, $1 \mapsto$ 1021, $2 \mapsto$ 2032, $3 \mapsto$ 3013. (d): states in $\mathbb{Z}_3 = \{0,1,2\}$ with local rule $\phi(x,y) = 2x + 2y$, initial configuration 11011011 and 9-substitution $0 \mapsto$ 000000000, $1 \mapsto$ 002020200, $2 \mapsto$ 001010100. (a) and (c) are rotationally symmetric, (b) and (d) are totally symmetric. The Hausdorff dimensions are respectively $\log 117/ \log 16 \approx 1.717$, $\log 36/ \log 8 \approx 1.723$, $\log 13/ \log 4 \approx 1.850$, $\log 54/ \log 9 \approx 1.815$.

Figure 2 shows a few examples of symmetric fractal limits obtained from rescaled cellular automata orbits generated by a proper initial top-row configuration subjected to a proper substitution as presented in the above theorem.

5 Conclusions

The above described method allows one to construct fractals which reflect the local rule of a cellular automaton as well as showing a global symmetry. Since these fractals are generated by a $(k \times k)$-substitution, these fractals can also be described in terms of hierarchical iterated function systems, in particular by so called k-adic HIFS[8]. By introducing a transition matrix[8] for the generating substitution it is possible to compute the fractal's box-counting dimension[5,6] which coincides with the Hausdorff dimension.

References

1. J.-P. Allouche, F. von Haeseler, E. Lange, A. Petersen, G. Skordev. Linear cellular automata and automatic sequences, *Parallel Computing* **23**(1997), 1577–1592.
2. Ch. Bandt. Self-similar sets I. Topological Markov chains and mixed self-similar sets, *Math. Nachr.* **142**(1989), 107–123.
3. A. Barbé. Symmetric coverings of triangles with Pascal's triangle modulo 2, SISTA/COSIC report nr. 98-54, Katholieke Universiteit Leuven, 1998 (submitted).
4. A. Barbé, F. von Haeseler. Cellular automata, quasigroups and symmetries, SISTA/COSIC report nr. 99-51, Katholieke Universiteit Leuven, 1999 (submitted).
5. G. Edgar. *Measure, Topology and Fractal Geometry*, Springer-Verlag, New York, 1990.
6. K. Falconer. *The Geometry of Fractal Sets*, Cambridge Univ. Press, New York, 1985.
7. F. von Haeseler, H.-O. Peitgen, G. Skordev. Pascal's triangle, dynamical systems and attractors, *Erg. Th. Dyn. Syst.* **12**(1992), 479–486.
8. F. von Haeseler, H.-O. Peitgen, G. Skordev. Linear Cellular Automata, Substitutions, Hierarchical Iterated Function Systems and Attractors, in *Fractals Geometry and Computer Graphics*, J.L. Encarnacao et al. (eds), Springer-Verlag, Heidelberg, 1992.
9. F. von Haeseler, H.-O. Peitgen, G. Skordev. Cellular automata, matrix substitutions and fractals, *Ann. Math. and Art. Intell.* **8** (1993), 345–362.
10. G. Hedlund. Endomorphisms and automorphisms of the shift dynamical systems, *Math. Syst. Theo.* **3**(4), 1969, 320–375.
11. R. Mauldin, S. Williams. Hausdorff dimension in graph directed construction, *Trans. Amer. Math. Soc.* **309** (1989), 811–829.
12. H.-O. Peitgen, H. Jürgens, D. Saupe. *Chaos and Fractals*, Springer-Verlag, New York, 1992.
13. S. Takahashi. Self-similarity of linear cellular automata, *J. Comput. Sci.* **44**(1992), 114-140.
14. S. Willson. Cellular automata can generate fractals, *Discr. Appl. Math.* **8**(1984), 91–99.
15. S. Wolfram. *Theory and application of cellular automata*, World Scientific Publ. Co. Pte. Lid., Singapore, 1996.

BISPECTRA AND PHASE CORRELATIONS FOR CHAOTIC DYNAMICAL SYSTEMS

ALLAN K. EVANS, STUART J. NIMMO AND MARK D. LONDON

Department of Mathematical Sciences, De Montfort University, The Gateway, Leicester LE1 9BH, UK. E-mail: ake@dmu.ac.uk

The bispectrum is the natural third-order generalisation of the power spectrum. It provides information about correlations between different Fourier components of a signal or image, and about the statistics of Fourier phase. A number of numerical and experimental studies of the bispectra of chaotic systems have been published. In this paper we present the first analytical calculations of the bispectra of chaotic dynamical systems. First, for a generalisation of the classical sawtooth or Renyi map, we calculate the bispectrum using symbolic dynamics. Also, for intermittent systems, we calculate the bispectrum using the relationship between these systems and renewal processes. We review the results of these calculations, drawing some conclusions about the characteristic features of the bispectra of chaotic systems, and compare them with the features of some financial time series.

Key words: Chaos, bispectrum, time series, intermittent, finance.

1 Introduction: power spectrum and phase in time-series analysis

Given an experimental time series x_t, what can we say about the nature of the system that produced it? This is a question commonly encountered by experimental scientists, and also by those who deal with financial time series such as stock prices or currency exchange rates. Often the first attack at the question is to calculate the power spectrum $P(\omega)$. The power spectrum reveals how the variance, or power, in the time series is shared between the different frequencies ω.

Because the power spectrum contains no phase information, it is not possible to reconstruct the original signal x_t from it, and there are some questions that can not be answered by inspecting only the power spectrum. One of these is the question of whether an apparently random signal originates from chaos in a low-dimensional deterministic dynamical system or from a stochastic process.[1,2]

The Fourier transform

$$\tilde{x}(\omega) = \sum_t x_t e^{i\omega t} \tag{1}$$

contains the same information as the original signal, and $\tilde{x}$ can be written in terms of a modulus and a phase:

$$\tilde{x}(\omega) = |\tilde{x}(\omega)| \, e^{i\phi(\omega)}, \tag{2}$$

where the modulus $|\tilde{x}(\omega)|$ determines the power spectrum. To go beyond the power spectrum while still using the concepts of Fourier analysis, one should therefore consider the statistics of the Fourier phase $\phi(\omega)$. In fact, the phase is more important in determining the appearance of an image than the Fourier modulus or power spectrum.[3] All information about the location of features in an image or the time of occurrence of features in a signal is contained in the phase.

It is not straightforward to study the statistics of phase in a time series. One advantage of the power spectrum is that it is a well-defined statistical average, being the Fourier transform of the autocorrelation function. However, stationary stochastic processes (including those defined by a chaotic dynamical system with its invariant measure) do not possess a well-defined average phase $\phi(\omega)$ or $e^{i\phi(\omega)}$ for any frequency ω.

We can understand why the Fourier phase factor $e^{i\phi(\omega)}$ does not have an average value for a stationary process by noting that translating the original time series by a delay s (that is, replacing x_t by $x_{(t-s)}$) is equivalent to multiplying each Fourier phase factor by $e^{i\omega s}$. This is just a restatement of the 'shift theorem' of Fourier analysis. Since any time-average of a stationary process must be invariant under a time-shift, this explains why the phase factor does not have a well-defined average.

This argument also provides a hint of how we might construct well-defined average quantities that contain phase information. For any two frequencies ω_1 and ω_2, the phase product

$$R(\omega_1, \omega_2) = e^{i\phi(\omega_1)} e^{i\phi(\omega_2)} e^{-i\phi(\omega_1+\omega_2)} \tag{3}$$

is invariant under a translation. A non-zero average of this quantity implies that the phases of the Fourier components at frequencies ω_1, ω_2 and $\omega_1 + \omega_2$ are correlated.

In fact, the mean value of R is the phase of the bispectrum,[4,5] evaluated at the frequencies (ω_1, ω_2). Just as the power spectrum $P(\omega)$ is the Fourier transform of the autocorrelation function, the bispectrum $P(\omega_1, \omega_2)$ is the double Fourier transform of the bi-correlation function

$$C_x(s, t) = \langle (x_s - \bar{x})(x_t - \bar{x})(x_0 - \bar{x}) \rangle, \tag{4}$$

where the angled brackets $\langle \rangle$ denote an expectation value.

The bispectrum can also be estimated directly from the Fourier transform:

$$P(\omega_1, \omega_2) = \lim_{T \to \infty} \frac{1}{2T} \langle \tilde{x}_T(\omega_1) \tilde{x}_T(\omega_2) \tilde{x}_T^*(\omega_1 + \omega_2) \rangle, \tag{5}$$

where the subscript T implies that the sum in equation (1) is over the first T samples in the time series, and $*$ denotes complex conjugation.

Higher-order generalisations of the power spectrum, known as polyspectra, can also be defined[4,5]. The bispectrum has been applied in time-series analysis to construct tests[6,7,8] for nonlinearity and for normality (meaning the presence of a normal distribution). These tests depend upon the observations that for normal systems, the bispectrum is zero, and for linear systems (that is, processes constructed by a linear operator acting upon an i.i.d. process) the bicoherence

$$b(\omega_1, \omega_2) = \frac{|P(\omega_1, \omega_2)|^2}{P(\omega_1) P(\omega_2) P(\omega_1 + \omega_2)} \tag{6}$$

is constant.

The bispectrum has also been used to study phase correlations and the interactions between Fourier components in many experimental systems, including electroencephalographic signals from the human brain,[9] and the oscillations of the earth.[10] A number of authors[8,11,12,13] have studied the bispectra of time series from chaotic dynamical systems, either by numerical simulation or experimental
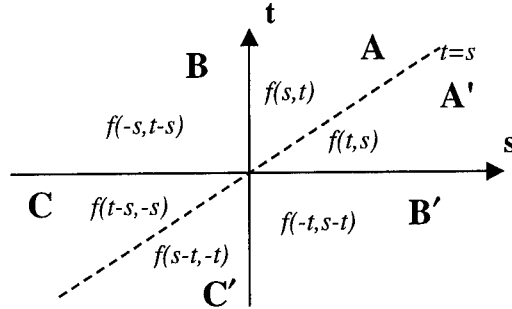
Figure 1. Symmetries of the bicorrelation function. For example, if the bicorrelation is given by $f(s,t)$ in the region A, then the bicorrelation in region C is given by $f(t-s,s)$.

measurement. Subba Rao[8] identified 'ridges' in the bicoherence as a characteristic of chaotic systems, and other authors studying electrical circuits have noted various changes in the bispectral characteristics of time series as the systems generating them make the transition to chaos.

To date, there are no published examples of chaotic dynamical systems where the bispectrum can be calculated exactly. In this paper we present an example of such a calculation, for a simple discrete-time system which is a generalisation of the well-known saw-tooth or Renyi map. We also show how the relationship between intermittent chaotic systems and renewal processes can be used to give asymptotic low-frequency expressions for the bispectra of intermittent systems. We identify some common features of these two families which may be bispectral characteristics of chaos. Finally, we compare some financial time series with the chaotic examples. Before presenting the calculations, we briefly review some properties of the bicorrelation function and bispectrum.

2 Symmetries of the bicorrelation function and calculation of the bispectrum

This section summarises some properties of the bicorrelation function and bispectrum that are useful in deriving the form of the bispectrum. From its definition in equation (4), one can see that, for a stationary process, the bicorrelation function $C_x(s,t)$ has the symmetries

$$C_x(s,t) = C_x(t,s) \qquad C_x(-s,t) = C_x(s,s+t). \tag{7}$$

By composing these symmetry operations, one can reach any part of the (s,t) plane from the region $0 < s < t$ where the bicorrelation is most easily calculated. These relationships are summarised in figure 1.

The bispectrum $P(\omega_1,\omega_2)$ is defined as the double Fourier transform of the bicorrelation function, the Fourier integral being evaluated over the entire (s,t) plane. Let $P_A(\omega_1,\omega_2)$ be the integral over only region A in figure 1. If the value of
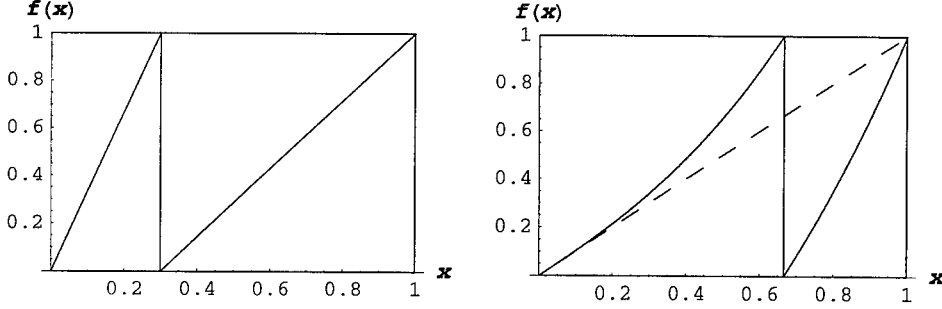
Figure 2. The two types of dynamical system considered in this paper: the generalised sawtooth map (left) with $b = 0.3$, and the intermittent dynamical system described by equation (28) with $\mu = 1$, $z = 2.7$.

the bicorrelation in region A is given by $C_x(s,t) = \beta(s,t)$, then

$$P_A(\omega_1, \omega_2) = \int_0^\infty ds \int_s^\infty dt \quad \beta(s,t) e^{i(s\omega_1 + t\omega_2)}. \tag{8}$$

Similarly, the integral over region B is

$$P_B(\omega_1, \omega_2) = \int_{-\infty}^0 ds \int_0^\infty dt \quad C_x(s,t) e^{i(s\omega_1 + t\omega_2)}. \tag{9}$$

Using the symmetries in equation (7), and changing the variables of integration, one can show that

$$P_B(\omega_1, \omega_2) = P_A(-(\omega_1 + \omega_2), \omega_2). \tag{10}$$

Similar relationships can be derived between P_A and the Fourier integrals over the other regions in figure 1. The bispectrum can then be calculated by summing all the integrals. The result is

$$P(\omega_1, \omega_2) = \hat{P}_A(\omega_1, \omega_2) + \hat{P}_A(-(\omega_1 + \omega_2), \omega_2) + \hat{P}_A(\omega_1, -(\omega_1 + \omega_2)), \tag{11}$$

where the hat operation $\hat{\cdot}$ denotes symmetrisation:

$$\hat{f}(x,y) = f(x,y) + f(y,x). \tag{12}$$

3 The generalised sawtooth map

3.1 The dynamical system

Our first dynamical system is defined by the mapping of the interval $(0,1)$ onto itself

$$f(x) = \begin{cases} \frac{x}{b} & x \le b \\ \frac{x-b}{1-b} & x > b \end{cases} \tag{13}$$

as pictured in Figure 2. The dynamical system $x_{t+1} = f(x_t)$ has a uniform invariant

distribution. It is chaotic, with Lyapounov exponent

$$\lambda = -b \ln b - (1 - b) \ln(1 - b) \tag{14}$$

easily calculated by standard methods.[14]

This dynamical system is a generalisation of the sawtooth or Renyi map which has been studied by many authors.[14] When $b = 1/2$, the two systems are identical. This generalisation makes the study of the system more difficult, but it is necessary in this case because when $b = 1/2$, the system's bispectrum is zero. This is an effect of the symmetry of the function f. Sakai and Tokumaru[15] have calculated the autocorrelation function of the $b = 1/2$ system using a relationship between this system and an autoregressive process.

3.2 Symbolic dynamics

The method we use to study this system is symbolic dynamics [14], where a sequence of symbols from a discrete and finite alphabet represents a trajectory of a dynamical system. The representation for the generalised sawtooth map is not difficult to derive. Suppose that x_0 and x_1 are two successive points in the trajectory of the system, so that $x_1 = f(x_0)$. Then we have either

$$x_0 = bx_1 \tag{15}$$

or, if $x_0 > b$,

$$x_0 = b + (1 - b)x_1. \tag{16}$$

Putting $z_1 = 0$ in the first case, and $z_1 = 1$ in the second, we can write

$$x_0 = z_1 b + x_1 e_1, \tag{17}$$

where $e_i = (b - 2z_i b + z_i)$. Continuing along the same lines, we obtain

$$x_0 = z_1 b + e_1(z_2 b + e_2(z_3 b + e_3 x_3))$$
$$= b e_1 e_2 \cdots e_{n-1} x_n + b \sum_{k=1}^{n} e_1 e_2 \cdots e_{k-1} z_k. \tag{18}$$

It is easy to show that, under the (uniform) invariant measure, the z_i are independent random variables, taking the values 0 or 1 with probabilities b and $(1 - b)$ respectively. We can therefore use this representation to calculate various averages. For example, in the next section we calculate the autocorrelation function and the power spectrum.

3.3 The power spectrum

Multiplying equation (18) by x_n, we obtain

$$x_0 x_n = b e_1 e_2 \cdots e_{n-1} x_n^2 + b \sum_{k=1}^{n} e_1 e_2 \cdots e_{k-1} z_k x_n. \tag{19}$$

We can now take the expectation value, using the fact that random variables with different subscripts are independent, and the expectation values $\langle z_i \rangle = (1 - b)$, $\langle z_i \rangle = (1 - b)$, $\langle x_i \rangle = 1/2$, $\langle x_i^2 \rangle = 1/3$ and

$$\langle e_i \rangle = b^2(1 - b)^2 = B, \tag{20}$$

$$\langle e_i^2 \rangle = b^3 + (1 - b)^3 = C, \tag{21}$$

$$\tag{22}$$

After summing a geometric series, we obtain

$$\langle x_0 x_n \rangle = \frac{1}{4} + \frac{B^n}{12}. \tag{23}$$

This equation is valid for $n \geq 0$. The autocorrelation function is therefore

$$C_x(t) = \langle x_0 x_t \rangle - \langle x_0^2 \rangle = \frac{B^{|t|}}{12}. \tag{24}$$

The power spectrum is the Fourier transform of this function (using the convention of equation (1)):

$$P(\omega) = \frac{1 + B^2}{12(1 + B^2 - 2B \cos \theta)}. \tag{25}$$

3.4 The bispectrum

The same methods can be used to calculate the three-point average $\langle x_0 x_s x_t \rangle$ when $0 \leq s \leq t$. We can then calculate the bicorrelation function. The details of this calculation occupy far too much space to be recorded here. The result is, again for $0 \leq s \leq t$,

$$C_x(s, t) = \frac{(1 - 2b)}{12} \left(B^t - B^s C^{t-s} \right), \tag{26}$$

where B and C are defined in equations (20) and (21). The bispectrum can now be calculated using the method described in the last section. The result is given by equation (11), with

$$P_A(x, y) = \frac{(2b - 1)(b - 1)be^{-ix}}{12 \left(1 - Be^{-ix} \right) \left(1 - Ce^{-ix} \right) \left(1 - Be^{iy} \right)} \tag{27}$$

Contour diagrams of the bispectrum and bicorrelation function for a typical value of b are shown in figure 3. The 'ridges' found by Subba Rao[8] are not evident for this system for any value of b that we have examined. The most noticeable consistent feature is that the energy in the bispectrum is spread over a broad ranges of frequencies rather than being concentrated into narrow bands. This may be a characteristic of fully-developed chaos.
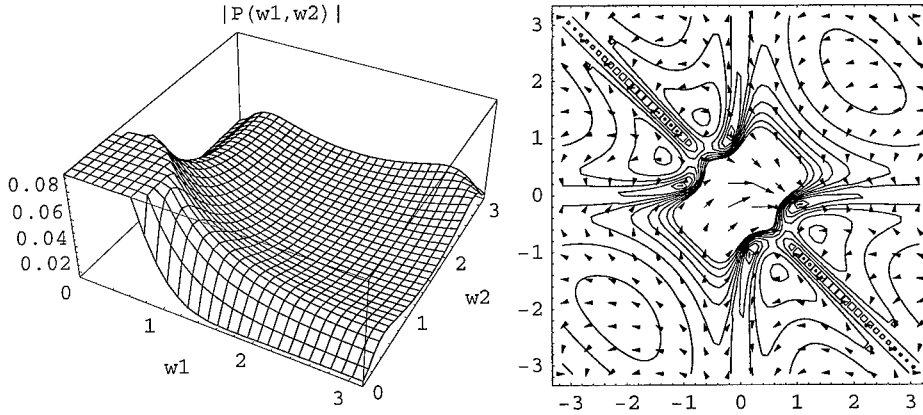
Figure 3. The bispectrum for the sawtooth map with $b = 0.27$. The left graph shows a surface map of the modulus of the bispectrum, evaluated over the region where ω_1 and ω_2 are positive. The right graph shows a direction field displaying the complex phase of the bispectrum overlaid with a contour map, evaluated over both positive and negative frequencies.

4 Intermittent dynamical systems

In this section, we study the bispectra of a class of intermittent dynamical systems. For definiteness, we take as our model the system defined by the mapping

$$f(x) = (x + \mu x^z) \bmod 1, \tag{28}$$

shown in figure 2 for $\mu = 1$, $z = 1.7$. The analysis given here depends only on the form of $f(x)$ close to the intermittent point at $x = 0$, and so applies to a broad class of intermittent dynamical systems,[14] as described by Ben-Mizrachi *et al.*[16] The defining characteristic of these intermittent systems is that they show brief periods of random behaviour between long periods when the system is 'stagnant', remaining close to the centre of intermittency (in the case of the model system above, close to the point $x = 0$).

We will be studying the long-time behaviour, or equivalently the low-frequency behaviour, of these dynamical systems. For this purpose, following Ben-Mizrachi *et al*,[16] we replace the dynamical system's time series x_t by a series of delta-functions located at the times of escape from periods of stagnation. The dynamical system is thus replaced by a renewal process with waiting-time distribution defined by the distribution of waiting times between escape events. As Ben-Mizrachi *et al*[16] show, this distribution can be approximated for large τ by

$$p_\tau(\tau) = \begin{cases} (\alpha - 1)t^{-\alpha} & t \geq 1 \\ 0 & t < 1, \end{cases} \tag{29}$$

where $\alpha = z/(z - 1)$. Ben-Mizrachi *et al* [16] also show that the function $c(t)$ giving the probability of an escape event at time t given the occurrence of such an event

at time 0 is given by

$$\bar{c}(s) = \frac{1}{1 - \bar{p}_\tau(s)}, \tag{30}$$

where the bar $\bar{\cdot}$ denotes the Laplace transform. The Laplace transform of p_τ is an incomplete Gamma function,[17] whose small-s behaviour is given by

$$1 - \bar{p}_\tau(s) \sim s^{\alpha-1}\Gamma(2 - \alpha) \tag{31}$$

for $1 < \alpha < 2$. For the sake of brevity, we consider only this case (corresponding to $z > 2$) in this paper. Other cases will be covered in a longer paper currently in preparation.

For the case $1 < \alpha < 2$, for small s we have $\bar{c}(s) \sim \Gamma(2 - \alpha)s^{1-\alpha}$, and the low-frequency behaviour of the power spectrum[16] is given by

$$P(\omega) = |\bar{c}(i\omega) + \bar{c}(-i\omega)|^2 \sim \frac{\text{constant}}{\omega^{2\alpha-2}}. \tag{32}$$

The bispectrum can also be calculated from $\bar{c}(s)$. First, note that for large t_1 and $(t_2 - t_1)$ and for $0 < t_1 < t_2$,

$$\langle x_{t_1} x_0 \rangle \sim c(\infty)c(t_1). \tag{33}$$

and

$$\langle x_{t_1} x_{t_2} x_0 \rangle \sim c(\infty)c(t_1)c(t_2 - t_1). \tag{34}$$

The bicorrelation $C_x(t_1, t_2)$ is therefore given by

$$C_x(t_1, t_2) = c(\infty)c(t_1)c(t_2 - t_1) - c(\infty)^2 \left(c(t_1) + c(t_2) + c(t_2 - t_1)\right) + 2c(\infty)^3. \tag{35}$$

After applying the double Laplace transform and taking the limit $s \to 0$, we find that the first term dominates and

$$\bar{C}_x(s_1, s_2) \sim \frac{c(\infty)}{\Gamma^2(2 - \alpha)s_1^{\alpha-1} (s_2 - s_1))^{\alpha-1}}. \tag{36}$$

By setting $s_1 = i\omega_1$, $s_2 = i\omega_2$, we obtain an expression for the bispectrum integral P_A:

$$P_A(\omega_1, \omega_2) \sim \frac{c(\infty)}{\Gamma^2(2 - \alpha) (i\omega_1 (\omega_2 - \omega_1))^{\alpha-1}}. \tag{37}$$

The bispectrum is then given by equation (11).

Bispectra calculated numerically from time series generated directly by the dynamical system (28) are in good agreement with this result, although a large number of data points is required for the average to converge to a smooth function. Figure 4 shows the result of one comparison.

The structure of these low-frequency bispectra is simple. They show power-law singularities at the lines $\omega_1 = 0$, $\omega_2 = 0$ and $\omega_1 + \omega_2 = 0$. Like the bispectra for the sawtooth map, the bispectra for intermittent systems (at least at low frequencies) show no isolated peaks. Intensity extends across a wide range of frequencies modulated by a power-law envelope.
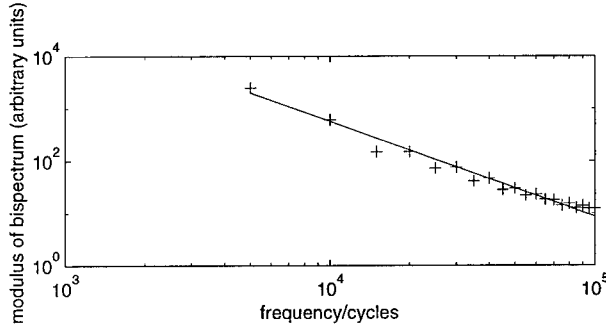
Figure 4. The theoretical treatment given in this paper predicts that the function $P(\omega_1, 0)$ should diverge as $\omega_1^{-2(1-\alpha)}$ as $\omega_1 \to 0$. This graph shows the theory (solid line) and values of a numerically calculated bispectrum using 10000 data points of a time series generated by the dynamical system of equation (28) for $z = 2.1$ ($\alpha = 1.909$).

These results suggest that bispectral measurements may be useful in cases where it is difficult to distinguish power-law noise from a dynamical system[16,2] from noise from a truly random source, such as filtered white noise. White noise retains its random phases when passed through a linear filter, and so has zero bispectrum.

5 Bispectra of financial time series

The bispectra of economic time series were first computed by Godfrey,[18] who was able to reject the null hypothesis of linearity for a number of stock price time series, but did not discuss other aspects of the bispectral form. For comparison with the other examples presented here, we calculated the bispectrum of the log-increments $y_n = \log(x_n/x_{n-1})$ of the Dow Jones Industrial Average stock market index, evaluated daily over the period January 1994 to February 1999. The bispectrum was estimated from the bicorrelation using a Parzen window, as in the work of Subba Rao and Gabr[7]. The result (figure 5) shows broad spectral intensity, as for the two chaotic examples presented here, but concentrated more at higher frequencies. This shift towards higher frequencies is probably a result of taking the log-increments, which is effectively a differentiation process.

6 Conclusion

One recognised characteristic of chaotic dynamical systems is that they show a continuous power spectrum, with intensity distributed across a wide range of frequencies. The results of this paper show that, at least in some cases, this is also true of their bispectra. This provides a way of distinguishing a chaotic time series from filtered white noise. However, other classes of processes share this characteristic of broad-band bispectral intensity. This is clear because the bispectra for the intermittent systems in this paper were calculated by exploiting the similarity

74
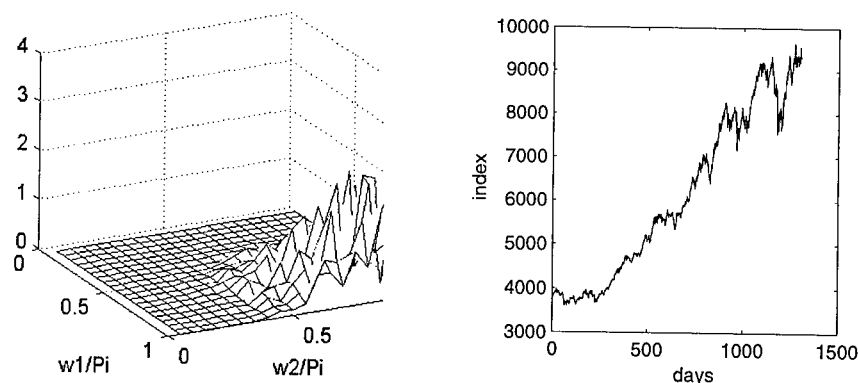
Figure 5. The left-hand graph shows the modulus of the bispectrum for the log-increments of the Dow Jones industrial average stock index (values shown on the right).

between these systems and renewal processes. A broad-band bispectrum can not therefore be taken as an unambiguous sign of chaos. A broad-band bispectrum does indicate interactions between Fourier modes over a wide range of frequencies, and is therefore a sign that the system studied does not have a simple description in the Fourier domain. The broad-band character of the bispectra of financial time series should therefore be taken as an indication of the complexity of the underlying processes.

References

1. P. Cvitanovic. *Universality in chaos.* Adam Hilger, Bristol, UK, 1984.
2. A.R. Osborne and A. Pastorello. *Phys. Lett. A,* 181:159–71, 1993.
3. A.V. Oppenheim. *Proc. IEEE,* 69:529–41, 1981.
4. M.B. Priestley. *Spectral analysis and time series.* Academic Press, New York, 1981.
5. J.M. Mendel. *Proc. IEEE,* pages 278–305, 1991.
6. M.J. Hinich. *Journal of time series analysis,* 3:169–176, 1982.
7. T. Subba Rao and M. Gabr. *Journal of Time Series Analysis,* 1:145–158, 1980.
8. T. Subba Rao. *Nonlinear modelling and forecasting (SFI studies in the sciences of complexity) M. Casdagli and S. Eubank (eds) Addison-Wesley,* 12:199–226, 1992.
9. J. Koren, L.J. Tick, R.A. Zeitlin, and C.T. Randt. *Bull. NY Acad. Med.,* 44:1127–1128, 1968.
10. M. Bozzi Zadro and M. Caputo. *Supplemento al nuovo cimento,* VI:67–81, 1968.
11. S. Elgar and V. Chandran. *IEEE trans. on circuits and systems 1: fundamental theory and applications,* 40:689–92, 1993.

12. K.B. Kim and S.Y. Kim. *J. Phys. Soc. Jap*, 65:2323–2332, 1996.

13. J.M. Lipton, K.P. Dabke, and M. Lakshaman. *Int. J. Bifurc. Chaos*, 6:2419–2425, 1996.

14. E.A. Jackson. *Perspectives of nonlinear dynamics (2 vols)*. Cambridge University Press, 1990.

15. H. Sakai and H. Tokumaru. *IEEE trans. on acoustics, speech and signal processing*, 28:588–590, 1980.

16. A. Ben-Mizrachi, I. Procaccia, N. Rosenberg, A. Schmidt, and H.G. Schuster. *Phys. Rev. A*, 31:1830–40, 1985.

17. M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions*. Dover, New York, 1965.

18. M.D. Godfrey. *Appl. Stat.*, 14:48–69, 1965.

SELF-ORGANIZED CRITICALITY MODELS OF NEURAL DEVELOPMENT

D.L. RAIL

Neurology Department, Campbelltown Hospital, Campbelltown NSW 2560, Australia
Email: drail@talent.com.au

B.I. HENRY AND S.D.WATT

Department of Applied Mathematics, School of Mathematics,
University of New South Wales, Sydney NSW 2052, Australia
E-mail: B.Henry@unsw.edu.au, S.Watt@unsw.edu.au

A simple evolutionary model is introduced for neural development along the lines of the Bak-Sneppen model for biological evolution of an ecology. The model represents a set of neurons and their connections together with associated synaptic weights. Evolution of the system is studied for different model fitness functions of the synaptic weights. The model systems exhibit Darwinian evolution of the synaptic weight space towards maturation.

1   Introduction

Until recently the biological contribution to learning was thought to be an unfolding according to an intrinsic schedule [1,2]. However, evidence now indicates that the developing cerebral cortex is largely free of domain-specific structure [3,4] and the representational properties of the cortex are built by the nature of the problems confronting it.

At the neurobiological level, learning stems from the interaction between intrinsic growth and environmentally derived activity [4]. Two factors that are fundamental for this interaction are selection and variance. Selection [5,6,7,8] has two distinct stages. The first stage constructs "pre-representations". The second stage selectively eliminates certain of these representations. The most fit representations survive to underlie mature skills.

Variation is important for the development of a maximally flexible representation capacity. The florid growth of neural tissues in ontogeny [4]: synapses; axonal and dendritic arborisation represents one possibility for introducing a significant chance element to learning. Representations are consolidated by a gradual increase in the synaptic weights of preferred units (synaptic weight space). The widely accepted mechanism for such consolidation is Hebbian learning via positive reinforcement. At the neural net level, Hebbian learning follows from the covariance of pre- and post-synaptic discharges. At the circuit level, it involves synergy of oscillators involved in common function. Synapses which do not participate in the maintenance of circuits are eliminated.

Selectivity and variability are also important aspects of biological evolution. One of the simplest models for biological evolution is the Bak-Sneppen model [9]. In this model evolutionary activity is simulated through random mutations of the least fit species and its neighbours. A characteristic feature of this model is that it exhibits Self-Organized Criticality whereby the system evolves through a succession

of punctuated equilibria to a state where almost all species have fitness above a threshold level.

In this paper we have introduced new variants of the Bak-Sneppen model to investigate the evolution of synaptic weight space towards maturation in simple models for neural development.

2 Neural Connectivity Model

Consider a set of n units (neurons) labelled $1, 2, 3, \ldots, n$ on a periodic one-dimensional lattice, i.e., units $k + 1$ and $k - 1$ are neighbours of unit k which is equivalent to unit $n + k$. Associated with each unit k we identify a set of $c(k)$ connections (synapses). Two versions of the models are studied. One in which the number of connections is kept fixed ($c(k) = m$) and the other in which the number of connections (for the least fit unit and its two neighbours) is selected at random, ($c(k) \in [1, m]$) at each update step. We have in mind that the former situation will be easier to investigate theoretically whereas the latter will be more representative of a model neural system. Let $c_j(k)$ denote the jth connection associated with unit k and identify a corresponding synaptic weight $a \le w_j(k) \le b$. Pre-representations are introduced by choosing the synaptic weight for each connection at random from a uniform distribution in the range $[a, b]$. Initially we suppose that the system is fully connected, i.e., initially $c(k) = m$ for all k. There is no discernible difference in the long term results whether or not the system is initially randomly connected. Selection and variation in the system is modelled by replacing the least fit unit and its two neighbours by new units with new randomly assigned connections and weights.

Three different models have been examined. In Model A the weights are random numbers in the range $[0, 1]$ and the fitness of a unit is defined as the minimum weight for that unit. In Model B the weights are random numbers in the range $[-1, 1]$ and the fitness of a unit is defined as the average of the weights for that unit. In Model C the weights are again random numbers in the range $[-1, 1]$ but the fitness is defined as the sum of the weights. The choice of the range $[-1, 1]$ in the latter two models is to include the possibility of both excitatory and inhibitory synapses. In each model the steps in the update procedure are as follows:

1. Set up the pre-representations which are fully connected with random synaptic weights.

2. Calculate the fitness for each unit and identify the unit with the lowest fitness.

3. For the least fit unit and each of its neighbouring units reset the number of connections depending on the variant of the model and assign new weights at random from a uniform distribution on $[a, b]$.

4. Return to step 2 and continue for N updates.

Model A with $m = 1$ is the Bak-Sneppen model. The transient behaviour of this model is characterized by punctuated equilibria as the system evolves towards a statistically stationary state in which the density of weights in the system with
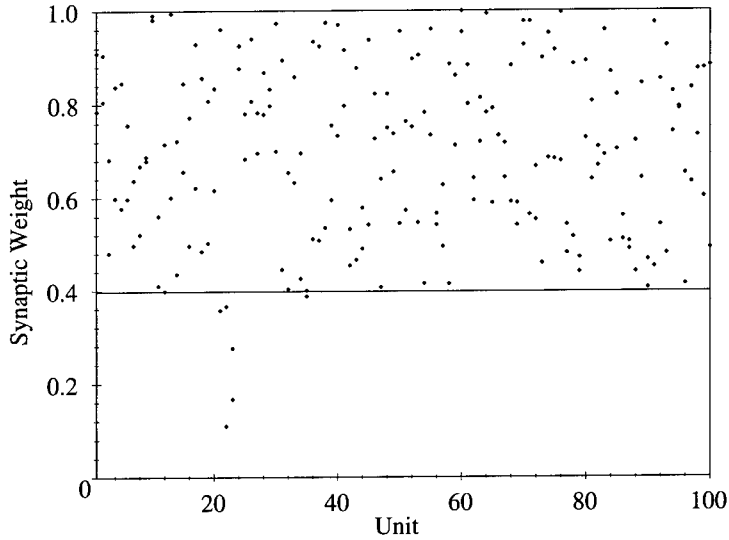
Figure 1. Snapshot of the stationary state for Model A with $m = 2$ and fixed numbers of connections.

$w < w^*$ vanishes. The critical weight $w^* \sim 0.66702 \pm 0.00003$ [10] is referred to as the self-organized threshold. A plot of the largest value of the minimum weight after s updates versus s reveals a staircase structure where the average length of a step in the staircase scales as a power law distribution. The change in the minimum weight across a level step in the staircase is referred to as an avalanche [9].

3 Simulations

Simulations of the models have been carried out for $n = 100$ units and $s = 10^6$ updates over a range of values of the maximum number of connections $m \in [1, 2^{10}]$; both for fixed number of connections and for random numbers of connections (up to the maximum m). To facilitate the discussion of these results let $f_m(s)$ denote the fitness of the least fit unit after the sth update in a model where m is the maximum number of connections.

In all cases the transient behaviour exhibits punctuated equilibria and the long term behaviour is highly correlated with almost all fitness values above a critical value. Figure 1 shows the weights for each unit plotted against the unit number in the 'stationary state', after 10^4 updates, in Model A with $m = 2$ and the number of connections fixed. Except for the localized avalanche, all weights are above the limiting threshold weight. Similar long term trends are found in all cases independent of m (compare for example, Figure 1 of [11] which shows a similar plot for the original Bak-Sneppen model), however the magnitude of the limiting threshold fitness is m dependent.

The transient behaviours with their characteristic punctuated equilibria are shown for each of the models in Figures 2a,2b,3a,3b,3c. These figures show plots of

(a)

(b)
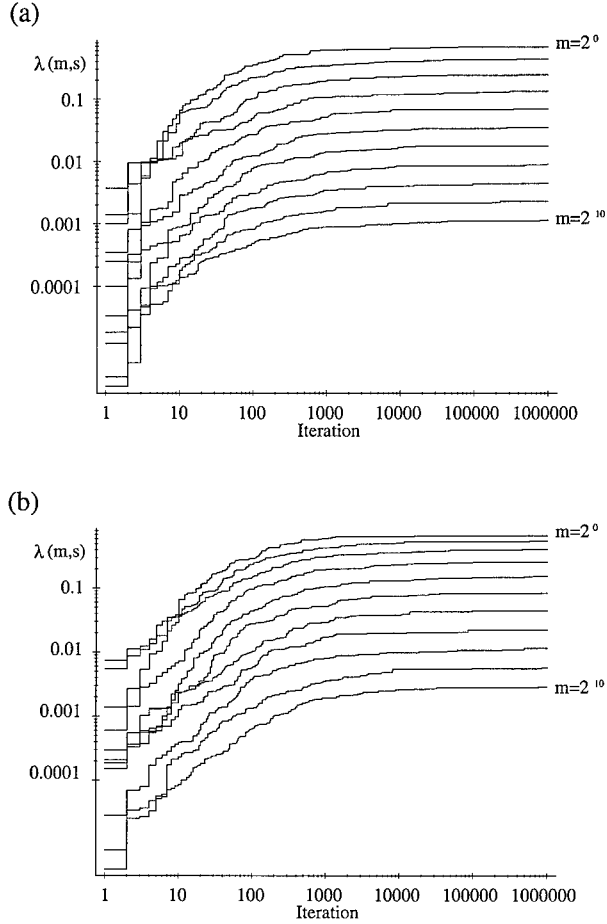
Figure 2. Devil's staircase structure for Model A; (a) fixed numbers of connections, (b) random numbers of connections.

the maximum value of the minimum fitness after s updates,

$$\lambda(m, s) = \max_{[0,s]} f_m(s),$$

versus s. In each horizontal step of the staircase structure in these figures the minimum fitness remains less than this threshold value. A rise in the staircase occurs when the minimum fitness exceeds this threshold value. The case of Model B with fixed numbers of connections was not simulated because this is the same as Model C with fixed numbers of connections apart from the scale factor m (see further comments below).

In Models A and B the limiting threshold fitness,

$$\lambda^*(m) = \lim_{s \to \infty} \max_{[0,s]} f_m(s),$$

decreases with an increase in the maximum number of connections m. In Model C
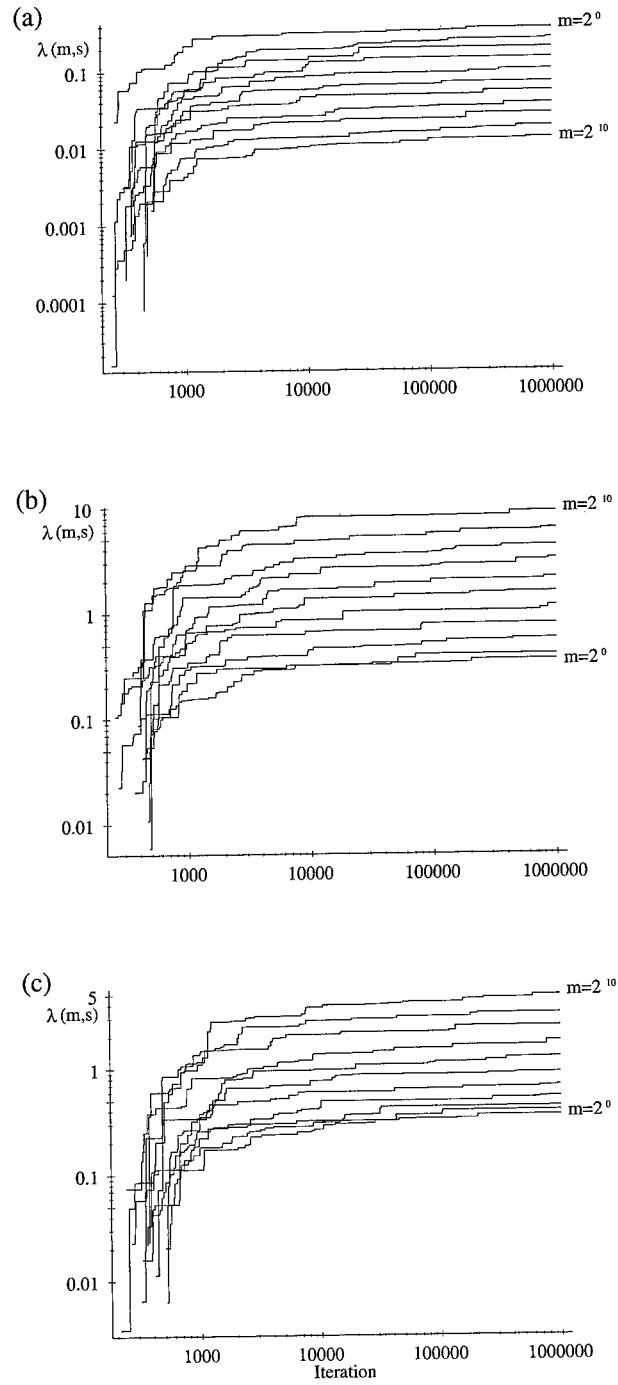
Figure 3. Devil's staircase structure; (a) Model B with random numbers of connections; (b) Model C with fixed numbers of connections, (c) Model C with random numbers of connections.
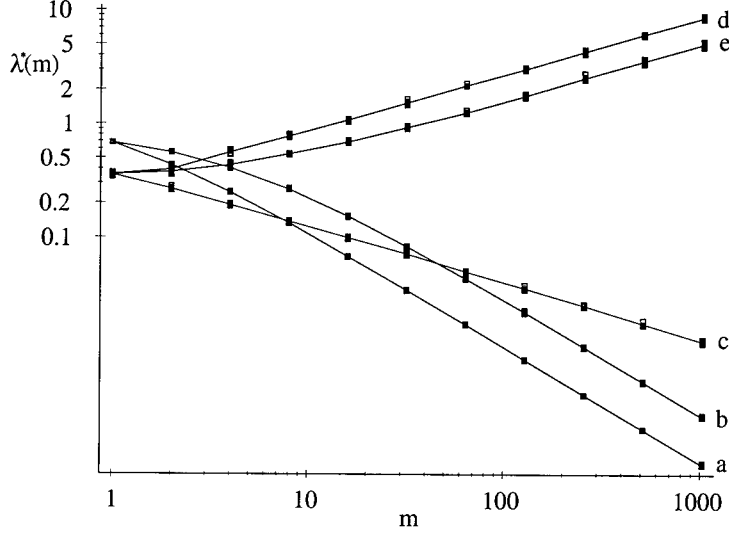
Figure 4. Plot of the limiting threshold fitness versus the maximum number of connections m: (a) Model A with fixed numbers of connections; (b) Model A with random numbers of connections; (c) Model B with random numbers of connections; (d) Model C with fixed numbers of connections; (e) Model C with random numbers of connections.

the opposite is true. Figure 4 shows plots of the limiting threshold fitness $\lambda^*(m)$ versus m for each of the models. The error bars in the plots are based on data from ten runs of each model and the lines are ensemble averages. From these plots we note that for large m: $\lambda^*(m) \sim \frac{1}{m}$ - for Model A; $\lambda^*(m) \sim \frac{1}{\sqrt{m}}$ - for Model B; and $\lambda^*(m) \sim \sqrt{m}$ - for Model C. For Model A the limiting value $\lambda^*(1)$ is the Bak-Sneppen self-organized criticality threshold value $0.66702\ldots$. For models B and C the limiting value $\lambda^*(1) \approx 2 \times 0.66702 - 1 = 0.33404$.

4 Scaling Analysis

The functional relationships between the limiting threshold fitness (the self-organized threshold) and the maximum numbers of connections can be obtained using simple probabilistic arguments. To this end (following [12]) we first define the avalanche probability function, $P_{\lambda(m)}(s)$ as the probability for an avalanche with threshold fitness $\lambda(m)$ which starts at $k = 0$ to end at $k = s$. It follows immediately from the definition that $P_{\lambda(m)}(0) = 0$. In Model A, the probability for the avalanche to end at step $k + 1$ is

$$P_{\lambda_A(m)}(k+1) = (1 - P_{\lambda_A(m)}(k))(1 - \lambda_A(m))^{3m}. \tag{1}$$

The first factor on the right hand side is the probability that the avalanche did not stop at the earlier step k and the factor $(1 - \lambda_A(m))^{3m}$ is the probability of independently selecting new fitness values at random from a uniform distribution for each of the $3m$ connections, for the least fit unit and its two neighbours, above

the threshold fitness $\lambda_A(m)$. It is a straightforward exercise to obtain the solution to the difference equation, Eq.(1). The solution for initial condition $P_{\lambda_A(m)}(0) = 0$ is

$$P_{\lambda_A(m)}(k) = \frac{(1 - \lambda_A(m))^{3m}}{1 + (1 - \lambda_A(m))^{3m}}(1 - (-1)^k(1 - \lambda_A(m))^{3mk}). \qquad (2)$$

At the limiting threshold values $\lambda_A^*(m)$ we anticipate that the avalanche probabilities $P_{\lambda_A^*(m)}(k)$ are independent of m. In particular by equating $P_{\lambda_A^*(m)}(k) = P_{\lambda_A^*(1)}(k)$ we obtain the relation

$$\lambda_A^*(m) = 1 - (1 - \lambda_A^*(1))^{\frac{1}{m}}, \qquad (3)$$

where $\lambda_A^*(1)$ is the limiting threshold fitness for the Bak-Sneppen model. The agreement between Eq. (3) and the values plotted in Figure 4 is summarized in Table 1.

We now consider the scaling relation for model C. In this case the avalanche probabilities are determined by the equation

$$P_{\lambda_C(m)}(k + 1) = (1 - P_{\lambda_C(m)}(k))(1 - \Phi\left(\frac{\lambda_C(m)}{\sqrt{m}\sigma}\right))^3 \qquad (4)$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{z^2}{2}\right) dz.$$

The second factor on the right hand side of Eq.(4) follows from the Central Limit Theorem. For sufficiently large m, the sum $Y(m)$ of m random numbers from a uniform distribution with zero mean and variance σ^2 is a Gaussian random variable Z with zero mean and variance $m\sigma^2$. Hence the probability of randomly selecting the sum $Y(m) \geq \lambda_C(m)$ is equal to the probability of selecting the Gaussian random variable $Z \geq \frac{\lambda_C(m)}{\sqrt{m}\sigma}$. This probability is given by

$$1 - \Phi\left(\frac{\lambda_C(m)}{\sqrt{m}\sigma}\right).$$

The solution of Eq.(4) for initial condition $P_{\lambda_C(m)}(0) = 0$ is

$$P_{\lambda_C(m)}(k) = \frac{(1 - \Phi\left(\frac{\lambda_C(m)}{\sqrt{m}\sigma}\right))^3}{1 + (1 - \Phi\left(\frac{\lambda_C(m)}{\sqrt{m}\sigma}\right))^3}(1 - (-1)^k(1 - \Phi\left(\frac{\lambda_C(m)}{\sqrt{m}\sigma}\right))^{3k}). \qquad (5)$$

The avalanche probabilities given by Eq.(5) are approximations based on the Central Limit Theorem which holds with increasing accuracy as m increases. The approximation is already reasonable for $m \geq 2$ but it does not hold in the case $m = 1$ where the sum over m random numbers from a uniform distribution on $[-1, 1]$ is simply the uniform random variable. In this case the probability for choosing a random variable greater than $\lambda_C(1)$ is $(1 - \lambda_C(1))/2$ so that the avalanche probability is given by

$$P_{\lambda_C(1)}(k) = \frac{(\frac{1 - \lambda_C(1)}{2})^3}{1 + (\frac{1 - \lambda_C(1)}{2})^3}(1 - (-1)^k(\frac{1 - \lambda_C(1)}{2})^{3k}). \qquad (6)$$

Table 1. Comparison between limiting threshold fitness values; $\tilde{\lambda}(m)$ obtained from the numerical simulations, and $\lambda(m)$ obtained from the theoretical scaling relations Eqs.(3,7). The percentage difference between these values is also shown.

m	$\lambda_A^*(m)$	$\tilde{\lambda}_A^*(m)$	% diff.	$\lambda_C^*(m)$	$\tilde{\lambda}_C^*(m)$	% diff.
1	0.67789	0.67789	0.00	0.35751	—	—
2	0.43259	0.43245	0.03	0.39351	0.39351	0.00
4	0.24769	0.24664	0.42	0.55863	0.55650	0.38
8	0.13330	0.13203	0.95	0.77354	0.78701	1.74
16	0.06815	0.06835	0.29	1.07521	1.11300	3.51
32	0.03501	0.03478	0.65	1.52210	1.57402	3.41
64	0.01774	0.01754	1.13	2.17175	2.22601	2.50
128	0.00884	0.00881	0.36	3.01230	3.14805	4.51
256	0.00442	0.00441	0.29	4.29448	4.45201	3.67
512	0.00222	0.00221	0.53	6.09033	6.29610	3.38
1024	0.00111	0.00110	0.47	8.49563	8.90402	4.81

The limiting threshold value for $m = 1$ in Model C is given by $\lambda_C^*(1) = 2\lambda_A^*(1) - 1$. To obtain the scaling behaviour with m of the limiting threshold values $\lambda_C^*(m)$ for $m \geq 2$ we now equate $P_{\lambda_C^*(m)}(k) = P_{\lambda_C^*(2)}(k)$ which yields

$$\lambda_C^*(m) = \sqrt{\frac{m}{2}}\lambda_C^*(2). \tag{7}$$

The good agreement between the $\lambda_C^*(m)$ values plotted in Figure 4 for $m \geq 2$ and Eq.(7) is summarized in Table 1.

The scaling behaviour in the case of model B can be obtained by repeating the analysis for model C but replacing $\Phi\left(\frac{\lambda_C(m)}{\sqrt{m}\sigma}\right)$ by $\Phi\left(\frac{\sqrt{m}\lambda_B(m)}{\sigma}\right)$ since the average of m random numbers from a uniform distribution with zero mean and variance σ^2 is a Gaussian random variable with zero mean and variance σ^2. This yields the scaling result

$$\lambda_B^*(m) = \sqrt{\frac{2}{m}}\lambda_B^*(2). \tag{8}$$

5 Discussion

The simple evolutionary models for neural development introduced in this paper attempt to model the chance aspect of learning via Darwinian evolution. All models exhibit Darwinian evolution of the synaptic weight space towards maturation where almost all neurons have fitness levels above a threshold value.

It is anticipated that the maximum number of connections scales with memory and learning. With this interpretation the 'fitness' functions in each of the models needs some clarification since from Figure 4 and the scaling analysis above we see that only Model C exhibits an increase of fitness with increasing memory and

learning. On this basis the fitness function in Model C is more representative of learning and memory than the fitness function in Models A and B. However a fitness function that scales with memory and learning can also be recovered from the present fitness function in models A and B simply by multiplying by an appropriate monotonically increasing function of m. For example multiply by m^γ where; $\gamma > 1$ for Model A, $\gamma > 1/2$ for Model B, and $\gamma > 0$ for Model C.

It is interesting that the scaling relations: $\lambda_A^*(m) \sim 1/m$, $\lambda_B^*(m) \sim 1/\sqrt{m}$ and $\lambda_C^*(m) \sim \sqrt{m}$ which were derived in the theoretical analysis (and obtained numerically) for the case of fixed numbers of connections are also obtained in our numerical simulations when the numbers of connections is selected at random (for the least fit unit and its neighbours) in each update. With the appropriate interpretation of fitness as above, for a fixed maximum number of connections, we find that the neurons self-organize themselves to operate at increasing fitness whilst at the same time decreasing the numbers of active connections. The scaling analysis of this reduction in the number of active connections with increasing fitness is deserving of further studies.

Towards the end of this study we became aware of the neuronal model of self-organized learning recently introduced by Chialvo and Bak [13]. Our models are similar to their model to the extent that memory and learning is consolidated via Darwinian elimination of the least fit units rather than via Hebbian re-inforcement of the most fit units. On the other hand our models are more crude than the Chialvo-Bak model since we do not include the dynamics of firing in our models.

References

1. E.H. Lenneberg, *Biological Foundations of Language* (Wiley, 1967).
2. N. Chomsky, *Behavioral and Brain Sciences* **3** 1 (1980).
3. W.T. Greenough, J.E. Black and C.S. Wallace, *Child Development* **58**, 539 (1987).
4. S. R. Quartz and T. J. Sejnowski, *Behavioral and Brain Sciences* **20**, 537 (1997).
5. N. Jerne in *The Neurosciences: A Study Program*, ed. G.C. Quarton, T. Melnechuk, F.O. Schmitt (Rockefeller University Press, 1967).
6. J.P. Changeux and A. Danchin, *Nature* **264**, 705 (1976).
7. P. Rakic, J. P. Bourgeois, M.F. Eckenhoff, N. Zecevic and P. S. Goldman-Rakic. *Science* **232**, 232 (1986).
8. G. M. Edelman, *Neural Darwinism. The Theory of Neuronal Group Selection* (Basic Books, 1987).
9. P. Bak and K. Sneppen, *Phys. Rev. Letts.* **71**, 4083 (1993).
10. M. Paczuski, S. Maslov, and P. Bak, *Phys. Rev. E* **53**, 414 (1996).
11. P. Bak and S. Boettcher, *Physica D* **107**, 143 (1997).
12. S. Boettcher and M. Paczuski, *Phys. Rev. Letts* **76**, 348 (1996).
13. D. Chialvo and P. Bak, *Neuroscience* **90**, 1137 (1999).

ALTERED FRACTAL AND IRREGULAR HEART RATE BEHAVIOR IN SICK FETUSES

MYUNG-KUL YUM

Department of Pediatric Cardiology, Hanyang University Kuri Hospital
249-1 Kyomun-dong Kuri-shi Kyunggi-do 471-701, Korea,
Email: mkyumm@email.hanyang.ac.kr

JUNG-HYE HWANG

Department of Obstetrics and Gynecology, Hanyang University Kuri Hospital
249-1 Kyomun-dong Kuri-shi Kyunggi-do 471-701, Korea,
Email: hwangjh@email.hanyang.ac.kr

MOON-IL PARK

Department of Obstetrics and Gynecology, Hanyang University School of Medicine,
17 Haengdang-dong Sungdong-ku, Seoul   133-792, Korea,
Email: parkmi@Email.hanyang.ac.kr

The purpose of this study was to show that abnormal fractal correlation and irregularity of heart rate behavior were altered in intrauterine growth restricted fetuses (IUGR group) and fetuses whose mothers had maternal pregnancy induced hypertension (PIH group). We analyzed fetal heart rate data of 5000 points in normal (n=98), IUGR (n=45), and PIH   (n=46) fetuses, with their gestational ages > 38 weeks and without any perinatal complication. We calculate approximate entropy for the quantifying irregularity, and short-range ($\leq$ 80 beats, α_1)   and long-range (>80 beats, α_2) fractal scaling exponent for quantifying the fractal correlation properties. We also performed spectral analysis. In the IUGR group, statistical, the spectral measures, α_2, and α_2/α_1 were significantly higher and the approximate entropy was significantly lower than in the normal group. In the PIH group, α_1 was significantly lower, and α_2 and α_2/α_1 were significantly higher. The fetuses associated with either IUGR or PIH, although they are not severely compromised, showed abnormal fractal and/or irregular heart rate behavior.

1.   Introduction

Antenatal fetal heart rate variability recordings have been widely used to evaluate fetal well-being.[1]   However, the conventional indices of heart rate variability such as mean, variance, short-term and long-term variability have shown limitations[2-3]   in providing information about the cardiovascular dynamics of healthy or diseased fetuses. Recently, newly developed nonlinear dynamical analysis of heart rate variability[4-6]   makes possible the uncovering of abnormal heart rate behaviors that are not apparent from using conventional measures. This technique is actively used in early detection of cardiovascular morbidity and prediction of mortality.[7-9] However, the application to find abnormal behavior of heart rates in sick fetuses was limited.

Of the new techniques, fractal correlation and irregularity measures of heart rate behavior arouse much interest.[8,10-11] The fractal measure quantifies the temporal correlation of each heart rate value with the previous one on short- and long-range scales, and thereby, smoothness or roughness of the landscape in the heart rate time series on each scale.[12]

The aim of present study was to test whether uncomplicated fetuses associated with intrauterine growth restriction (IUGR) and pregnancy induced hypertension (PIH), which are the most frequently encountered fetal sicknesses, had abnormalities in fractal and irregular heart rate behavior.

2. Methods

2.1. Subjects

Among the pregnant women who had visited the outpatient obstetric clinic of Hanyang University Hospital, 98 uncomplicated cases of pregnant women and 45 cases of pregnant women whose fetuses were suspected of IUGR and were confirmed after delivery were, included in this study. We also included the 46 fetuses whose mothers had severe PIH. Severe PIH was defined when one or more of the following criteria were satisfied: Blood pressure $\geq$ 160/110 mmHg, proteinuria $\geq$ 2+ on reagent strip, oliguria < 400mL/24 hours, elevated liver enzymes, platelet < 100,000/mm^3, and pulmonary edema. All were single pregnancies, with gestational ages between 39 and 42 weeks which were calculated using the date of the last menstrual period and early ultrasound measures. All births were without any major malformations, chromosomal anomalies, and perinatal complication. Using growth standards of Brenner et al.[13], birth weights of infants were determined against gestational ages. Infants with birth weights below the 10th percentile were regarded as IUGR. At the time of recording, pregnant women in labor or taking drugs, including smoking, which can affect the fetal heart rate variability, were excluded.

2.2. Data collection

All subjects were in a semirecumbent position for a minimum of 10 minutes before data collection. Fetal heart rate tracings were recorded for more than 40 minutes using a Corometrics 115 (Corometrics, Model 115, USA) external fetal monitor. The recorded data were sampled into a personal computer with a digital serial interface. Whenever missing data were found, they were recorded as zero. Above 10 percent signal loss in the entire data have been excluded. When the off-line fetal heart rate data of zero (missing data) or below 30 beats per minute or above 200 beats per minute were encountered, they were removed. We selected 5000 points-data which corresponded to about 30 minutes of recording during which fetal movements were included.

2.3. Power spectral analysis

Power spectral analysis was performed by classical fast Fourier transformation.[14] We calculated low-, and high-frequency power by integrating the power spectral density curve at area between 0.03 and 0.15 Hz, and 0.15 and 0.4 Hz, respectively.[11]

2.4. Detrended fluctuation analysis

To quantify the fractal correlation behavior, we employed the detrended fluctuation analysis,[4,8,15] which is a modification of classical root-mean square analysis of a random walk. The heart rate time series (length N=5000) was first integrated,

$$y(k) = \sum_{i=1}^{k}[x(i) - x_{ave}]$$

where x(i)is the ith heart rate and x_{ave} is the average heart rate. Next the integrated time series was divided into boxes of equal length, n. In each box of length n, a least squares line was fit to the data. The y-coordinate of the straight-line segments is denoted by $y_n(k)$. Next, we detrended the integrated time series, y(k), by subtracting the local trend, $y_n(k)$, in each box. The root-mean-square fluctuation of this integrated and detrended time series is calculated by

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k)-y_n(k)]^2}$$

This computation was repeated over all time scales (box sizes) to provide a relationship between F(n) and n. Typically F(n) will increase with n. A linear relationship on a double-log graph indicates the presence of scaling. The slope of the line relating log F(n) to log (n) determines the fractal scaling exponent. However, in most of the fetuses' heart rate time series, we found that log-log plot was not strictly linear but rather consisted of two distinct linear regions of different slopes separating at a break point near 80 beats (Fig.1, Crossover point).
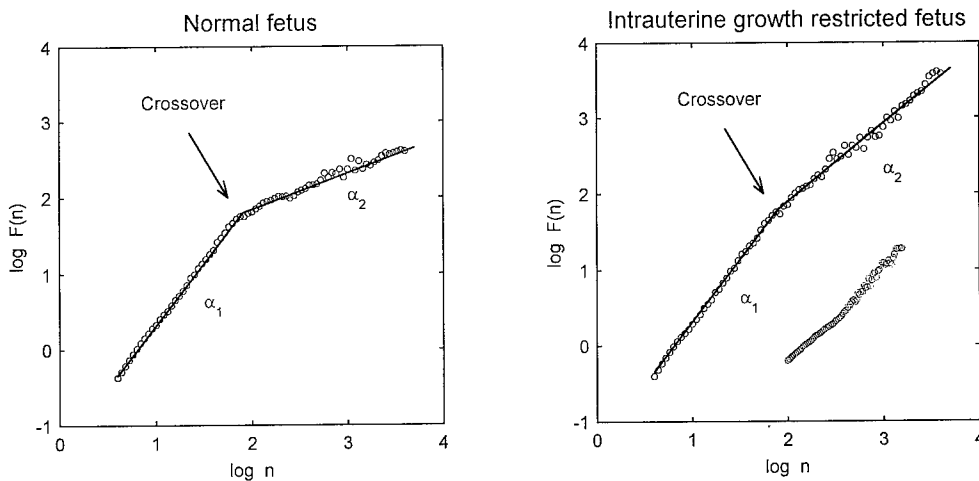
Figure 1. Plots of log F(n) vs log (n) (see text) in a normal and a growth restricted fetus aged 40 weeks of gestation. For both fetuses, detrended fluctuation analysis curves are approximated linear over two regions, with a slope α_1 for small values of n ($\leq$80 heartbeats, short-range fractal scaling exponent) and α_2 for large values of n (>80 heartbeats, long-range scaling exponent), resulting in apparent crossover phenomena. Note α_1 is similar in both fetuses (1.253 vs 1.248). However, α_2 is significantly higher in the growth restricted fetus (1.092 vs 0.652). Arrows indicate crossover points.

Therefore, fractal correlation of heart rate was defined separately for short-range ($\leq$ 80 beats, short-range fractal scaling exponent, α_1) and long-range (> 80 beats, long-range fractal scaling exponent, α_2) fluctuation of heart rate (Fig.1). Crossover index was defined as α_2/α_1.

2.5. *Approximate entropy.*

The methodological details for computing approximate entropy, ApEn(m,r,N), have been published elsewhere[5,8,11,16-17] and will be described briefly. In order to compute ApEn(m,r,N), three input parameters should be fixed (m : the length of compared runs, r : the effective filter, N : the length of data points). Let each heartbeat interval data set be represented as x(i).

ApEn(m,r,N)=average over i of $\log_e$[conditional probability that |x(j+m)-x(i+m)|≤r, given that |x(j+k)-x(i+k)|≤ for k=0,1,..., m-1]. (1)

From x(i), vector sequences u(1) through u(N-m+1) are formed, defined by u(i)=[x(i),...,x(i+m-1)]. These vectors represent m consecutive x values, commencing with the ith point. Define the distance d[u(i), u(j)] between vector u(i) and u(j) as the maximum difference in their respective scalar components. Use the sequence u(1), u(2),...., u(N-m+1) to construct, for each i≤ N-m+1, $C_i^m(r)$= (numbers of j ≤ N-m+1 such that d[u(i), u(j)]≤ r)/(N-m+1). Define

$$\Phi^m(r) = \sum_{i=1}^{N-m+1} \frac{\log_e C_i^m(r)}{N-m+1}$$

, and then define the approximate entropy ApEn(m,r,N)

$$= -[\Phi^{m+1}(r) - \Phi^m(r)] = -[\sum_{i=1}^{N-m} \frac{\log_e C_i^{m+1}(r)}{N-m} - \sum_{i=1}^{N-m+1} \frac{\log_e C_i^m(r)}{N-m+1}]$$

$$\cong -\sum_{i=1}^{N-m} \frac{\log_e C_i^{m+1}(r) - \log_e C_i^m(r)}{N-m} = (N \to \infty) - \sum_{i=1}^{N-m} \frac{\log_e[C_i^{m+1}(r)/C_i^m(r)]}{N-m}$$

which equals the average over i of

$$\log_e[C_i^{m+1}(r)/C_i^m(r)]$$

This last parenthetical expression is readily seen to be the conditional probability indicated in Eq. (1). For this study N = 5000 points and the two parameters of standard deviation of each data set were m=2 and r=20% standard deviation of each data set. Fig. 2 plots all conditional probabilities over i=1,2,...,5000,

$$C_i^{m+1}(r)/C_i^m$$

of two heart rate data in an normal and an IUGR fetuses. Approximate entropy was the negative of the average logarithmic conditional probability, i.e.,

$$-\sum_{i=1}^{N-m} \frac{\log_e[C_i^{m+1}(r)/C_i^m(r)]}{N-m}$$
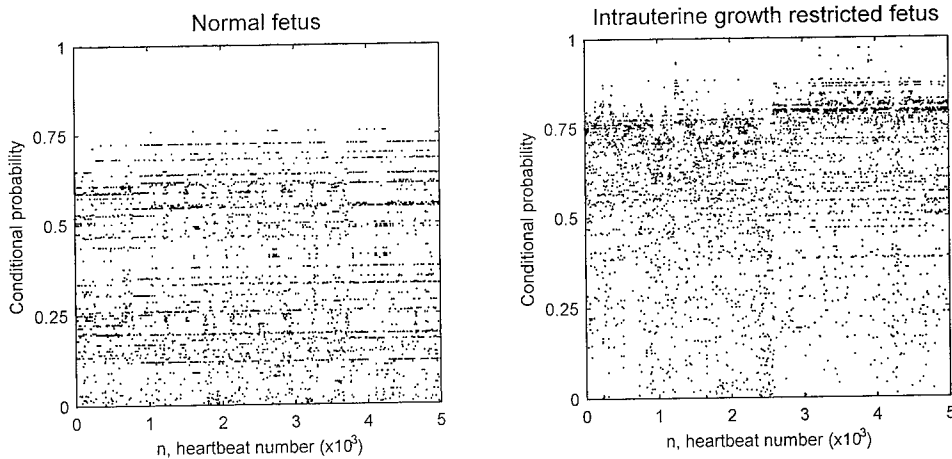
Figiure 2. Plots of all conditional probabilities for calculating the approximate entropy of the two fetuses shown in Fig.1, which is the negative of the average logarithmic conditional probabilities. The approximate entropy is much less in the growth restricted fetus than in the normal fetus (1.043 versus 0.532).

2.6. Statistical analysis.

All data are presented as mean ± 1 standard error of mean. To test the statistical difference in the involved parameters between the three groups, a procedure for general linear model (Statistical Analysis System 6.11) was used. When a significant statistical difference in the parameter among the three groups was identified, the difference between the two groups was compared with Duncan test. All analyses were performed using an alpha level of 0.05 as the criterion for statistical significance.

3. Results

In the IUGR group, the mean birth weight was 2426 g ± 30, compared to 3267 g ± 22 in control group and 3134 g ± 22 in PIH group ($p<0.05$, $p<0.05$). The results of statistical, spectral and nonlinear analysis are summarized in table 1.

In the growth restricted group, the mean and the variance of heart rate were significantly higher compared to those of normal and PIH group. The frequency domain measures, low- and high-frequency power, were also significantly higher. However, the approximate entropy was significantly lower. Detrended fluctuation analysis revealed that all three groups showed apparent crossovers exhibited for the fractal scaling behavior of heart rate. The emergence of the crossover phenomena was due to a higher short-range fractal scaling exponent (α_1) compared to a long-range fractal scaling exponents (α_2) and, therefore, crossover index (α_2/α_1) was far less than unity. The coefficient α_1 of the IUGR group was not significantly different from that of the normal group. However, the value of α_2 and α_2/α_1 were significantly higher than those of the normal group. All indexes except the fractal measures in the PIH group were not significantly different to those of normal group. $\alpha1$ was significantly lower, and α_2 and α_2/α_1 were significantly higher.

Table 1. Statistical, spectral, and nonlinear indexes of heart rate in normal and IUGR group for entire gestational periods.
Data are represented as mean ± standard error of mean. IUGR: uncomplicated intrauterine growth restriction; PIH: pregnancy induced hypertension; bpm: beats per minute; * <0.05, normal vs IUGR; †p<0.05, normal vs PIH; ‡p<0.05, IUGR vs PIH.

	Normal group (n = 98)	IUGR group (n = 45)	PIH group (n = 46)
Mean (bpm)	141.7 ± 0.8	145.3 ± 1.4*‡	142.4 ± 1.2
Variance (bpm^2)	47.2 ± 3.5	85.6 ± 9.2*‡	47.8 ± 6.4
Low-frequency power (msec^2)	132.8 ±9.5	204.3 ± 29.5*‡	113.1 ± 1.0
High-frequency power (msec^2)	23.8 ± 1.6	33.5 ± 5.0*‡	22.9 ± 2.1
Approximate entropy	0.687 ± 0.017	0.578 ± 0.023*‡	0.704 ± 0.028
Short-range scaling exponent (α_1)	1.451 ± 0.107	1.419 ± 0.122	1.401 ± 0.015†
Long-range scaling exponent (α_2)	0.713 ± 0.022	0.970 ± 0.028*	0.915 ± 0.027†
Crossover index (α_2/α_1)	0.492 ± 0.015	0.689 ± 0.023*	0.660 ± 0.022†

4. Discussion

By demonstrating the higher long-range fractal scaling exponent and crossover index in the IUGR and PIH group without any perinatal complication, and lower short-range fractal scaling exponent in the PIH group, it was uncovered that, although they were not severely compromised, their fractal heart rate behaviors were abnormal. By showing the decrease in approximate entropy in the IUGR group, it was also revealed that their irregular heart rate behavior was altered. In the PIH group, unlike in the IUGR group, only fractal measures could differentiate them from normal group. Furthermore, to differentiate the two groups from normal one, only the long-range fractal scale exponent or crossover index was useful.

The approximate entropy quantifies the irregularity in the time series by measuring randomness and nonlinearity in it.[5,16-17] Therefore, one distinguishing feature of the heart rate behaviors in the IUGR fetuses is that each heart rate tends to overshoot less or undershoot less (decreased nonlinearity) and to change less in random directions (decreased randomness).

The fractal scaling exponent measures absolute degree of temporal correlation and smoothness or roughness in heart rate fluctuation.[4,8,12]

Although the average value of short-range fractal scaling exponent in the PIH group was significantly higher, the values of all three groups were near 1.5. Therefore, on the short-range scale, their heart rate resembles smooth Brownian noise. The average value of the long-range fractal scaling exponent in the normal group was 0.713, which is near 0.5, and over long-range scale, the landscape of the heart rate time series resembles rough white noise, whereas the value in the IUGR and PIH group were near unity and their landscape resembles 1/f noise. Therefore, a distinguishing feature of their heart rate behavior from that of the normal group is that on the long-range scale the heart rate fluctuation is more timely correlated and smoother. The increased low-frequency oscillation evidenced by increased low-frequency power in the IUGR group (Table 1) and their possible increase in the duration of heart rate acceleration that occurred during fetal movement, fetuses might account for the increased long-range correlation and smoothness. However, this is not the case, since the long-range fractal scaling behavior operates at 80 - 1000 heartbeats, whose duration is beyond the period of the low-frequency and the period of the acceleration.

The abnormal fractal correlation behavior and/or decreased irregularity in the heart rate of the IUGR and PIH group may reflect the abnormalities in integrated complex cardiovascular control,[8,18] which may impair the fetuses' ability to adapt to external and internal perturbations and predispose the fetuses to perinatal mortality and morbidity.[19-21] This notion is supported by recent observations that either decreased long-range fractal scaling exponent[7,8-10] or approximate entropy[11,22] in heart rate dynamics was associated with patients with a variety of cardiovascular diseases.

In conclusion, by demonstrating abnormal fractal scaling measures and/or decreased approximate entropy in fetuses associated with IUGR and PIH and without any perinatal complications, we found that, although they are not severely compromised, their fractal and irregular heart rate behaviors were altered.

References

1. Farmakides G, Weiner Z. Computerized analysis of the fetal heart rate. *Clin. Obstet. Gynecol.* 38(1995) pp. 112-120.
2. Gagnon R., Bocking A.D., Richardson B.S. and McLean P., A flat fetal heart rate tracing with normal fetal heart rate variability. *Am. J. Obstet. Gynecol.* 171 (1994) pp. 1379-1381.
3. Samueloff A., Langer O., Berkus M., Field N., Xenakis E. and Ridgway, L., Is fetal heart rate variability a good predictor of fetal outcome? *Acta. Obstet. Gynecol. Scand.* 73 (1994) pp. 39-44
4. Peng C.K., Mietus J., Hausdorff J.M., Havlin S., Stanley H.E. and Goldberger A.L.. Long-range anticorrelations and non-Gaussian behavior of heartbeat. *Phys. Rev. lett.* 70 (1993) pp. 1343-1346
5. Pincus S.M. and Goldberger A.L., Physiological time-series analysis: what does regularity quantify? *Am. J. Physiol.* 266 (1994) pp. H1643-H1656
6. Mestivier D., Chau N.P., Chanudet X., Bauduceau B. and Larroque P., Relationship between diabetic autonomic dysfunction and heart rate variability assessed by recurrence plot. *Am. J. Physiol.* 272 (1997) pp. H1094-9

94

7. Makikallio T.H., Ristimae T., Airaksinen K.E.J., Peng C.K., Goldberger A.L. and Huikuri H.V., Heart rate dynamics in patients with stable angina pectoris and utility of fractal and complexity measures. *Am. J. Cardiol.* **81** (1998) pp. 27-31

8. Ho K.K.L., Moody G.B., Peng C.K., Mietus J.E., Larson M.G., Levy D. and Goldberger A.L., Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation* **96** (1997) pp. 842-848

9. Nelson J.C., Rizwan-uddin, Griffin M.P. and Moorman J.R., Probing the order within neonatal heart rate variability. *Pediatr. Res.* **43** (1998) pp. 823-31

10. Iyengar N., Peng C.K., Morin R., Goldberger A.L. and Lipsitz L.A., Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *Am. J. Physiol.* **271** (1996) pp. R1078-R1084

11. Hogue C.W. Jr., Domitrovich P.P., Stein P.K., Despotis G.D. Re L., Schuessler R.B., Kleiger R.E. and Rottman J.N., RR interval dynamics before atrial fibrillation in patients after coronary artery bypass graft surgery. *Circulation* **98** (1998) pp. 429-34

12. Peng C.K., Havlin S., Stanley H.E. and Goldberger A.L., Qunatification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos* **5** (1995) pp. 82-87

13. Brenner W.E., Edelman D.A. and Hendricks C.H., A standard of fetal growth for the United States of America. *Am J Obstet Gynecol* **126** (1976)   pp. 555-62

14. Akselrod S., Gordon D., Ubel F.A., Shannon D.C., Barger A.C. and Cohen R.J., Power spectrum analysis of heart rate fluctuation: a qunatitative probe of beat to beat cardiovascular control. *Science* **213** (1981) pp. 220-222

15. Hausdorff J.M., Mitchell S.L., Firtion R., Peng C.K., Cudkowicz M.E., Wei J.Y. and Goldberger A.L., Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease. *J. Appl. Physiol.* **82** (1997) pp. 262-269

16. Pincus S.M. and Viscarello R.R., Approximate entropy: A regularity measure for fetal heart rate analysis. *Obstet. Gynecol.* **79** (1992) pp. 249-55

17. Pincus S.M. and Singer B.H.,   Randomness and degrees of irregularity. *Proc. Natl. Acad. Sci. USA.* **93** (1996) pp. 2083-2088

18. Elbert T., Ray W.J., Kowalik Z.J., Skinner J.E., Graf K.E. and Nirbaumer N., Chaos and physiology: deterministic chaos in excitable cell assemblies. *Physiol. Rev.* **74** (1994) pp. 1-47

19. Galbraith R.S., Karchmar E.J., Piercy W.N. and Low J.A., The clinical prediction of intrauterine growth retardation . *Am. J. Obstet. Gynecol.* **133** (1979) pp. 281-6

20. Williams R.L., Creasy R.K., Cunningham G.C., Hawes W.E., Norris F.D. and Tashiro M.   Fetal growth and perinatal viability in California. *Obstet. Gynecol.* **59** (1982) pp. 624-32

21. Low J.A., Galbraith R.S., Muir D., Killen H., Pater B. and Karchmar J., Intrauterine growth retardation:a study of long-term morbidity. *Am. J. Obstet. Gynecol.* **142** (1982) pp. 670-7

22. Pincus S.M., Cummins T.R. and Haddad G., Heart rate control in normal and aborted SIDS infants. *Am. J. Physiol.* **264** (1993) pp. R638-646

EXTRACT MULTIPLE SCALING IN LONG-TERM HEART RATE VARIABILITY

D.C. LIN

Department of Mechanical Engineering
Ryerson Polytechnic University
Toronto, Ontario, Canada, M5B 2K3
E-mail: derlin@acs.ryerson.ca

R.L. HUGHSON

Department of Kinesiology
University of Waterloo
Waterloo, Ontario, Canada, N2L 3G1
E-mail: hughson@healthy.uwaterloo.ca

Many natural processes can be characterized by their scale-invariance property. In this study, we present the results of potential multiple scalings in the long-term heart rate data from young healthy adults subjected to normal daily activity. Our approach is based on the direct check of the probabilistic structure of the increment process. Results from fractional Brownian motion are compared and the generating mechanism for multiple scaling is discussed in the context of scale-invariance formalism.

Keywords: scale invariance, multiple scaling, heart rate variability

1 Introduction

Many physical processes are scale invariant, giving rise to renormalizable structure and self-similarity in space and time. It is quite often that the renormalization can be conducted at different scales with different scaling exponents, namely, a multiple scaling property. Multiple scaling was observed in many natural phenomena ranging from physical[1,2,3,4,5,6], biological[7,8,9,10], to economical systems[1,11,12]. When a multiplicative mechanism is involved in the dynamics, *multifractal* theory can be used to characterize a (continuous) set of singularity exponents and the (Hausdorff) dimension of their supports via a Legendre transformation. Successful applications of the formalism revealed deep insights about many natural processes such as the density profile in diffusion limited aggregates[13,14], the velocity and dissipation fields in fully developed turbulence[15,2,3], the money exchange index from financial market[1,11,12] and network traffic[5,6]. In general, there is not an unified theory for multiple scaling and, sometimes, only finite number of scaling exponents can be ascertained. In this work, we will present such an example in the long-term heart rate variability (HRV) from healthy young adults and discuss its generating mechanism.

The study of HRV has continued to draw great interests in recent years. It is of both fundamental and clinical importance. In particular, a "healthy" heart typ-

ically shows a power-law like power spectrum[16], which can imply scale invariance[a] and "self-similarity" in the autonomous nervous system. Losing such variability was found to correlate well with the mortality rate of heart diseased patients[17]. The variation in the HRV scaling has been discussed by a number of researchers in the past[7,8,10]. In particular, Peng et. al. developed the systematic method and compared the power-law scalings in the very short time scale (a few heart beats) with the "asymptotic" behaviour (above 1000 heart beats). Significant difference was concluded in these widely separated time scales and the characteristic was found sufficient to distinguish between the healthy subjects and patients with congestive heart failure[10]. Di Rienzo et. al. also noted the variation of the power-law exponent of HRV in frequency, which in turns causes variations of the power-law scaling of the blood pressure as well[7]. Later, Viswanathan et. al. tackled the non-uniformity of power law scaling between the healthy subjects and those with severe heart disease from a point-process aspect[18]. Again, the conclusion was drawn on the distinguishability between the health and disease. In these previous studies, multiple scaling within physiologically relevant range of healthy humans has not been explored.

In this work, multiple scaling was studied based on the data increment process. The increment process at various time lags enables us to focus on the *local scale invariance* property of the data. The numerical method was developed and tested on artificial time series with one- and multiple- scaling exponents and then applied to the heart rate data set. We also compared with the existing methodology in the literature and found noticeable difference in the result. This paper is organized as follows. The main ideas and the numerical method are introduced in the next section. The results of artificial time series and the evidence of multiple scaling in heart rate data are presented in section 3. In the last section, the mechanism generating the multiple scaling in HRV will be discussed in the context of scale invariance formalism.

2   Extract Scale Invariance from the Probability Density Function

In this section, we will first recall the basic definition of scale invariance of a process and then describe the numerical method to extract multiple scaling. Given a time series, $\mathbf{r}(s)$, its scale invariance is defined by the family of probability density functions of the increment, $\Delta \mathbf{r}(t; s) = \mathbf{r}(t+s) - \mathbf{r}(s)$: for any λ and a fixed constant h, one has

$$f_{\lambda t}(n) = \lambda^{-h} f_t(\lambda^{-h} n) \tag{1}$$

where f_t denotes the probability density function of $n \equiv \Delta \mathbf{r}(t; s)$ and the constant h is the scaling index or scaling exponent of $\mathbf{r}(s)$.

To study (1) numerically requires the assumption of stationarity. This appears to hold in our application. When the increment is stationary, $f_t(n)$ can then be estimated from the ensemble of $\{\Delta \mathbf{r}(t; s), s = 1, 2, \cdots\}$. Due to the data fluctuation, it was found more efficient to fit $f_t(n)$ with a specific form, say $g_t(n)$, than

---

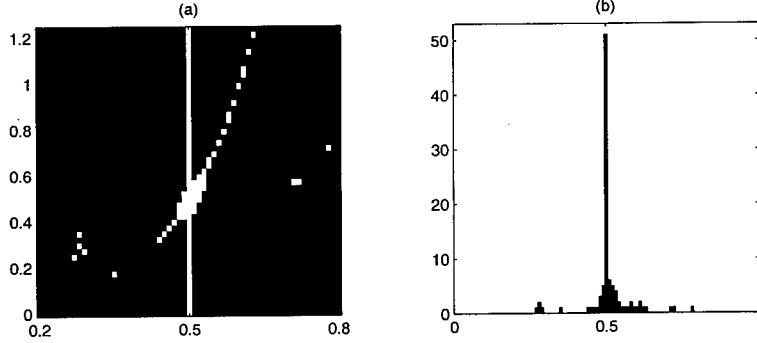[a]Note that power-law spectrum is only a necessary condition[19,20].

Figure 1. (a) The set $\mathcal{N}(\mathbf{T};h)$ plotted on the $h - n$ plane for $\mathbf{T} = \{2^k,\ k = 0,1,2,3\}$ and $h_o = 0.5$, dark pixels: $|\mathcal{N}| = 0$, white pixels: $|\mathcal{N}| > 0$, (b) $\mu(\mathbf{T};h)$ vs. h plot. The family of density functions used in this demonstration was ideal in that they have the Gaussian form: $\{g_t(n) = \exp(-n^2/2\sigma(t)^2)/\sqrt{2\pi\sigma(t)^2}, \sigma(t) = \sqrt{t};\ t = 2^k, k = 0,\cdots 11\}$.

working directly with the histogram of $\Delta\mathbf{r}(t)$ in (1). For example, for the fractional Brownian motion, a Gaussian form is assumed for $g_t(n)$. In this case, the maximum likelihood method was used to minimize the likelihood function $-\log(\Pi_i g_t(n_i))$ in order to extract the mean and variance of a Gaussian probability density function. This approach is subjected to less bias comparing to, say, minimizing the L^2 norm between the histogram and $g_t(n)$[21]. After $g_t(n)$ is defined, (1) can be studied by systematically varying n, λ and h_o values. Let

$$u(n,\lambda;h) = \frac{f_{\lambda t}(n)}{f_t(\lambda^{-h}n)}. \tag{2}$$

and denote the estimated slope, $d\log(u)/d\log(\lambda)$, as h'. $\mathbf{r}$ is said to be renormalizable or is scale invariant with a scaling exponent h_o if $h' \sim h_o$.

The parameters used in the numerical experiment were $t = 1$, $\lambda \in \{2^k, k = 0,\cdots,11\}$ and $h \in \{0.01 \cdot k,\ k = 0,\cdots,100\}$. The estimated slope h' is in general a function of n, h, and λ. The range of λ in which h' is estimated reveals the time scale of the local scale invariance property of $\mathbf{r}(s)$. Once the parameters are defined, we first construct, for a given h, $\mathcal{N}(\mathbf{T};h) = \{n,\ |h'(n,\lambda;h) - h| < \epsilon\}$ with $\epsilon = 0.01$ to keep the error between h' and h less less than 1%. Here, $\mathbf{T}$ denotes the set of λ's where the condition $|h' - h| < \epsilon$ is satisfied. For any given h and $\mathbf{T}$, $\mathcal{N}$ contains the set of n's in which local scale invariance is defined. The scaling interval of h' can thus be calculated by the logarithm of the ratio of the largest and smallest λ in $\mathbf{T}$. For each h, (1) also contains an isolated solution (when the set $\mathcal{N}$ has only one element) which has no relevance in the current context (Fig. 1a). Fortunately, these isolated solutions form a small set. Hence, given any $\mathbf{T}$ and h, $\mu(\mathbf{T};h) = |\mathcal{N}(\mathbf{T};h)|$, where $|\cdot|$ returns the number of elements at a specific h value, will show a peak at the desired scaling exponent (Fig. 1b).
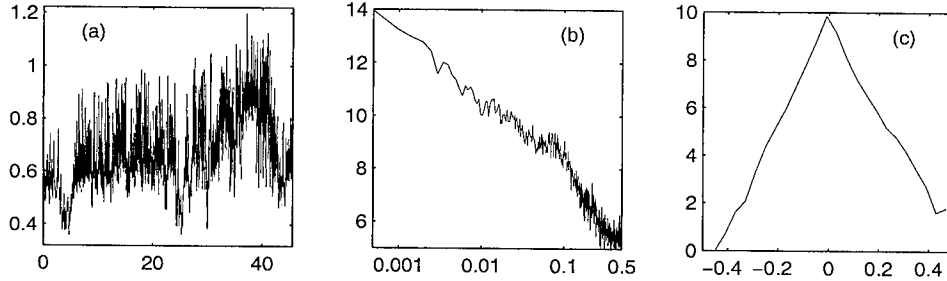
Figure 2. (a) Typical RR-interval record ($\mathbf{r}(s)$ (sec) vs. s ($\times 10^3$ beat)), (b) the power spectral density function of $\mathbf{r}(s)$ (in log-log scale). (c) histogram of $\Delta\mathbf{r}(t), t = 4$ (in linear-log scale).

3 Numerical Evidence of Multiple Scaling in HRV

3.1 Experimental Procedure

Six young adults (average age: 25 yr, height: 173 cm, weight: 74 kg) participated in the experiment. The subjects were allowed to conduct their normal daily activity. The difference in the body surface potential was sampled at 1000 Hz for a period of approximately 24 hours. The data was then down-loaded to a PC and a specialized computer code was written to search for the QRS complex for each heart beat with proper filtering for events such as skip-beat, PVC, and so on. The time span between the successive contractions of the ventricles, measured as the RR-interval (RRi), was finally extracted and used in the scale invariance analysis.

A representative day-time RRi data is shown in Fig. 2a. Scale invariance in the RRi data may be implied by the power-law trend of the power spectrum (Fig. 2b). But it is clear that a single power-law is not sufficient to describe the spectrum over the full frequency range.

3.2 Numerical Results

Before applying the numerical method to the RRi data, we first tested it on the fractional Brownian motion of scaling exponent $h_o = 0.26, 0.5, 0.78$. Each artificial time series had 40,000 samples and was generated by using the spectral method[19,20]. Fig. 3a shows a typical case of the set $\mathcal{N}$ plotted on the $h - T$ plane where $\mathbf{T} = 2^k, k = 1, 2, \cdots$ Sometimes, it is useful to consider $\sum \mu(\mathbf{T}; h)$ over all $\mathbf{T}$ sets for a given h. The result is shown in Fig. 3b where the scaling exponent is indicated at the location of the peak. It is clear that the desired scaling exponents were captured with good accuracy.

We next apply the spectral method to construct a two-exponent artificial time series of 2^{16} data points (Fig. 4a). The local scale invariance in the time series was defined by the scaling exponent $h_o = 0.26$ for frequency below 0.0025Hz and $h_o = 0.78$ above 0.0025Hz. Fig. 4d was plotted with $\sum \mu(\mathbf{T}; h)$ versus h and the estimated h's were within 3% of the exact values. In Fig. 4c, we checked the minimum time scale of the exponents, which is defined by the smallest element of
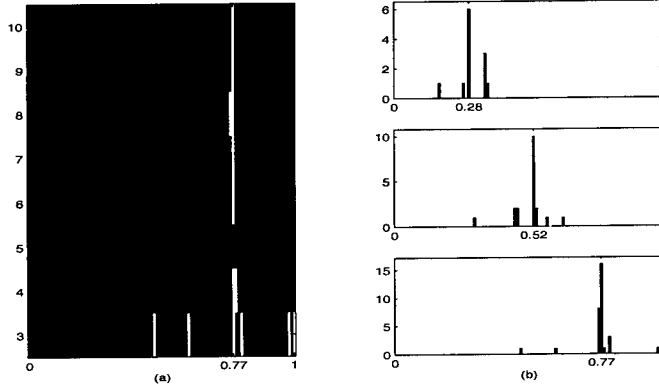
Figure 3. (a) The set μ on the $h - \mathbf{T}$ plane, dark pixels: $|\mu| = 0$, white pixels: $|\mu| > 0$ for $h_o = 0.78$.
(b) $\sum \mu(\mathbf{T}; h)$ vs. h for $h_o = 0.26, 0.5, 0.78$.

the $\mathbf{T}$ set. This is obtained by constructing the set $\eta(T; h)$ where T is the smallest λ in the corresponding $\mathbf{T}$ set. The set η shown on the $h - \log_2(T)$ plane reveals the minimum time scale from which the exponent is defined (Fig. 4c). For the artificial time series, it was found that correct time scales of the exponent were captured: i.e., the "fast" dynamics (above 0.0025Hz) of $h_o = 0.78$ prevails in small λ's and the "slow" dynamics of $h_o = 0.26$ at large λ (see Fig. 4c caption).

To study the RRi data, the following form of probability density function was assumed (based on Fig. 1c):

$$g_t(n) = \frac{\exp(-|n|^\alpha/t^\beta)}{\gamma}. \tag{4}$$

The RRi data of all the test subjects exhibit multiple scaling characteristics. In what follows, we will present the evidence from a typical individual whose data has been given in Fig. 2. In Fig. 5a, multiple peaks were seen from the $\sum \mu$ vs. h plot, indicating multiple scaling of HRV. At least five "significant" exponents were identified ($a \sim e$ in Fig. 5a). The minimum time scale of these exponents scattered over the range of $\lambda = 2^1$ to 2^8 (Fig. 5b). The inverse of this range covers the "frequency" < 0.0039 (1/beat) to 0.5 (1/beat) in the power spectrum (Fig. 2b). In this range, two linear regions of slopes $\zeta \sim -1.4$ and -2.1 may be roughly defined. Based on $\zeta = 1 + 2h$, they match nicely to two of the peak locations $h = 0.17$ and 0.6 in Fig. 5a. The size of the scaling interval is shown by the set $\mathcal{N}$ plotted on the $h - |\mathbf{T}|$ plane (Fig. 5c). A wide range of scaling interval associated with the scaling exponents found in Fig. 5a is again seen.

Finally, we compared our method with the *detrend fluctuation analysis* (DFA)[8,10] and found noticeable difference. In Fig. 6 shows the double-logarithmic plot of the average variance versus the scaled window length. Only one exponent can be ascertained here since the local scale invariance characteristics have been averaged out in the process of DFA.
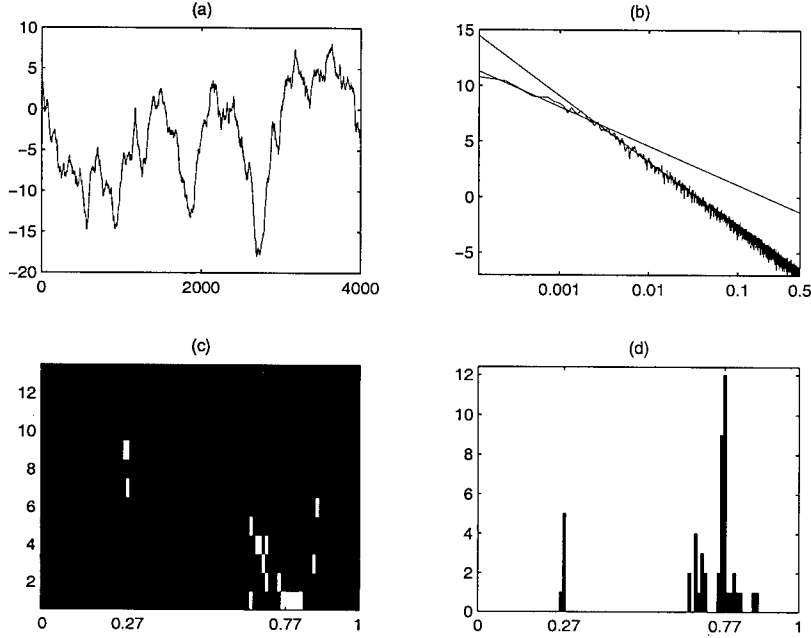
Figure 4. (a) Two-time-scale artificial time series, (b) power spectral density of the time series: the scaling exponents are $h_o = 0.26$ for $f < 0.0025Hz$ and $h_o = 0.78$ for $f > 0.0025Hz$. (Note: $0.0025 = 2^{-8.6}$). The line segments have slops of -1.52 and -2.56, respectively. (c) The set η plotted on $h - \log_2(T)$ plane, dark pixels: $|\eta| = 0$, white pixels: $|\eta| > 0$. The scaling range for $h = 0.77$ was captured for $T < 2^7$ and that for $h = 0.27$ for $T = 2^7 \sim 2^9$. (d) $\sum \mu(\mathbf{T}; h)$ vs. h plot.
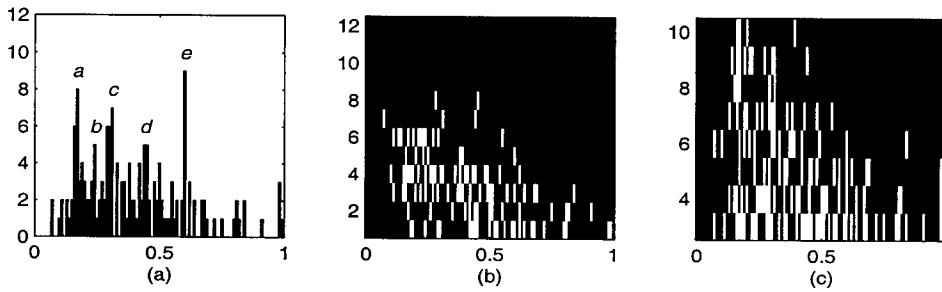
Figure 5. (a) $\sum \mu(\mathbf{T}; h)$ vs. h plot. Identified peaks at h=0.17, 0.24, 0.31, 0.44, 0.6 are labeled as a, b, c, d, e, respectively; (b) the set η plotted on $h - \log(T)$ plane, dark pixels: $|\eta| = 0$ and white pixels: $|\eta| > 0$, (c) The set of $\mathcal{N}$ plotted on $h - \log_2(T)$ plane, dark pixels: $|\mathcal{N}| = 0$, white pixels: $|\mathcal{N}| > 0$.

4 Discussion and Future Outlook

Numerical evidence of multiple scaling of HRV in physiologically relevant range has been presented for the case of healthy young adults. The scale invariance was
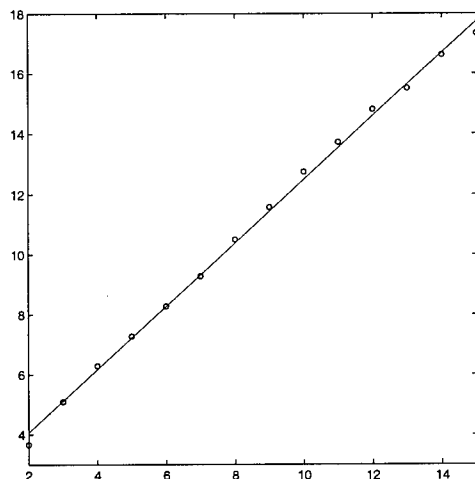
Figure 6. Detrend fluctuation analysis on the RRi data shown in Fig. 2a: averaged variance of the detrend time series vs. window length plot on double logarithmic scales (symbol: 'o'). The fitted line of slope ~ 1.05, suggesting $h = 0.05$, was shown as the solid line.

extracted based on the property of the family of RRi increment probability density functions. This approach enables us to explore local scale invariance property in the data. Although it is only a bit costlier in computation, details previously unavailable from other global methods, such as the power spectral density function or DFA, may be obtained. It should be noted that the current methodology is not able to extract temporal structure of very short duration. When such a rare event occurs, its characteristics will yield to that generated by the more "regular" dynamics in the process of estimating the density function. What this work has shown is that even the "regular" neuro-control of our cardiovascular system is rich enough to encompass a wide range of time scales and scaling structures, i.e., multiple scaling. Indeed, the process is very complex, as Hausdorff and Peng showed from artificial time series that the "balance" of different inputs to the system is also crucial to generate the $1/f$ scaling[8].

More detailed characterization of the density function (4) is underway and will provide the insight of the generating mechanism of multiple scaling. For example, the variation of the parameters α, β and γ in (4) can imply self-similarity of the individual density function at specific time lag (λt). This in turns can lead to scale invariant solution of (1) over finite n-interval. In particular, given a $\Delta \in \mathbf{R}$ and $n' \in I_n(\lambda t)$, assume the graph $f_{\lambda t}(n)$ is self-similar in the interval I_n:

$$\lambda^\Delta f_{\lambda t}(\lambda^\Delta n') = f_{\lambda t}(n'). \tag{5}$$

Since it can be written

$$f_{\lambda t}(n') = f_{\lambda t}(\lambda^h n''), \tag{6}$$

(5) implies

$$\lambda^\Delta f_{\lambda t}(\lambda^\Delta n') = \lambda^{-h} f_t(n'') \tag{7}$$

where $n' = \lambda^h n''$. Re-arranging terms in (7), one finds

$$\lambda^{h+\Delta} f_{\lambda t}(\lambda^{h+\Delta} n'') = f_t(n'') \tag{8}$$

which is of the same form as (1). Hence, (8) implies the existence of a new exponent $h + \Delta$. Substituting the newly found exponent into the scale invariance formalism (1), even more can be revealed by following the same arguments. In general, we found, when certain conditions are met, the number of exponents can "grow" as a power-law by repeating (5) with all the exponents. However, the scaling interval I_n also shrinks at the same time by a power-law, making most of the exponents "unobservable[22]" . We will report more details in a future publication.

Acknowledgments

References

1. A. Hilgers and C. Beck, *Int. J. Bif. & Chaos* **7**, 1855 (1997).
2. V.S. L'vov and I. Procaccia, *Phys. Rev. Lett.* **76**, 3963 (1996); *Phys. Rev. Lett.* **77**, 3541 (1996); *Phys. Rev.* E **54**, 6268 (1996).
3. B.B. Mandelbrot, *J. Fluid Mech.* **62**, 331 (1974).
4. Y. Pesin and H. Weiss, *J. Stat. Phys.* **86**, 233 (1997).
5. R.H. Riedi, M.S. Crouse, V.J. Ribeiro and R.G. Baraniuk, *IEEE. Trans. Info. on Theo.* **45**, 992 (1999).
6. J.M. Tchéou, M.E. Brachet, F. Belin, P. Tabeling and H. Willaime, *Physica* D **129**, 93 (1999).
7. M. Di Rienzo, G. Parati, A. Pedotti and P. Castiglioni, *Frontiers of Blood Pressure and Heart Rate Analysis*, Ed. M. Di Rienzo, G. Mancia, G. Parati, A. Pedotti and A. Zanchetti (IOS Press, 1997).
8. J.M. Hausdorff and C.-K. Peng, *Phys. Rev.* E **54**, 2154 (1996).
9. D.C. Lin and K Behdinan, *Proc. 1st Canadian Conference on Nonlinear Solid Mechanics*, **2**, 589 (1999).
10. C.K. Peng, S. Havlin, H.E. Stanley and A.L. Goldberger, *Chaos, Soliton & Fractal* **5**, 82 (1995).
11. B.B. Mandelbrot, A. Fisher and L. Calvet, *Cowles Foundation*, Paper No. 1164 (1997).
12. A. Fisher, L. Calvet and B.B. Mandelbrot, *Cowles Foundation*, Paper No. 1166 (1997), http://www.econ.yale.edu/ fisher/papers.html
13. C. Amitrano, A. Coniglio, P. Meakin and M. Zannetti, PRB **44**, 4974 (1991).
14. A. Coniglio, *Physica* A **140**, 51 (1986).
15. U. Frisch, *Turbulence: The Legacy of A.N. Kolmogorov* (Camb. Press. 1995).
16. M. Kobayashi and T. Musha, *IEEE. Trans. on Biomed. Eng.* **29**, 456 (1982).
17. J.T. Bigger Jr., J.L. Fleiss, R.C. Steinman, L.M. Rolnitzky, R.E. Kleiger and J.N. Rottman, *Circulation* **85**, 164 (1992).

18. G.M. Viswanathan, C.-K. Peng, H.E. Stanley and A.L. Goldberger, *Phys. Rev. E* **55**, 845 (1997).
19. N.P. Greis and H.S. Greenside, *Phys. Rev. A* **44**, 2324 (1991).
20. T. Highchi, *Physica* D **31**, 277 (1988); *Physica* D **46**, 254 (1990).
21. S. Mercik and R. Weron, *Physica* A **267**, 239 (1999).
22. D.C. Lin and R.L. Hughson, in preparation.

A SEMI-CONTINUOUS BOX COUNTING METHOD FOR FRACTAL DIMENSION MEASUREMENT OF SHORT SINGLE DIMENSION TEMPORAL SIGNALS-Preliminary study

V. PEAN, M. OUAYOUN, C.H. CHOUARD, B. MEYER

Ear, Nose and Throat Research Laboratory
Centre-Hospitalo-Universitaire Saint-Antoine
184 rue du Fbg Saint-Antoine; F75012 - PARIS - France
E-mail: <vincent.pean>,<michel.ouayoun>,<claude-henri.chouard>,<bernard.meyer>@sat.ap-hop-paris.fr

Box counting method allows to measure the eventual fractal dimension (D) of a single dimension temporal signal. However its accuracy varies as a function of the frequency sampling (Fs) and the duration of the tested signal (Sd). Consequently, as it is impossible to highly increase Fs, this method is not suitable for short physical signals D measurement. Thus, we designed a semi-continuous box counting method (SCBC) allowing a better approach of the small scales of the signal, especially useful in case of short single dimension temporal signal.

Let N = number of samples of the tested signal. SCBC provides with the first M points of the graph log – log owing to the dyadic division of boxes at large scales up to a certain box size S_M, such as $S_M = 2^M/Fs$. Then, at smaller scales, for each successive point the box size decreases by 1/Fs, that provides the – log with a large number of points. Thus, when $N/S_{(M+x)}/Fs$ is not a whole number, the analyzed signal is peripherally and symmetrically reduced in abscissa and ordinate, so that a whole number of boxes is obtained. But these truncated samples are then reintroduced for designing following boxes. Using SCBC we measured D of mathematical signals which D is known, and compared these results to those obtained using the classic dyadic box counting method.

1 Introduction

A continuous physical signal such as a sound signal, of which speech is an example, constitutes a time series. If this series is self-similar, measurement of its fractal dimension (D) allows fluctuations to be characterised by quantifying the complexity and irregularity of this signal. The classic box counting method seems the most appropriate to quantify complexity in the temporal succession of events. As the box size is divided by 2 for each measurement, this method may be termed dyadic box counting method (DyBC). In the framework of our research to improve signal processing in auditory prosthesis [1,2,3], we believed that this eventual D of speech elements could help patients with implants to recognize speech without having to lip-read.

In a recent study [10] we carried out a fractal approach of vowels. But, unlike mathematical signals, a physical signal cannot be fractal at every scale. However, let be N(r) the number of boxes filled by events at the resolution r, by definition the generalised dimension D_{gen} of the graph of the signal amplitude vs time is obtained using the limit

$$D_{gen} = \lim_{r \to 0} [\log(N(r))] / [\log(1/r)]$$

Thus, in our study, we had to use the higher resolution to approach infinitely small time scales, and to estimate the eventual fractality of vowels. In this aim we could observe that it is possible to appreciate, at least, the tendency of the points set by calculating the slope of the last 3 points, i. e. D of small size boxes (ssD). We demonstrated that:

- in case of sinusoidal signal for which $D_{gen} = 1$, ssD allows a better approach of the value of D_{gen} than using all points for the slope calculation;
- in the particular case of speech, ssD is able to supply a kind of significant signature of the vowels signal.

However we noted that the accuracy of DyBC varies as the ratio of [frequency sampling / frequency of the tested signal]. But we observed also that this accuracy depends on the duration of the tested signal. Indeed several authors, e. g. Robinson [5], estimate that at least 10 points are necessary to characterise the eventual fractality of a physical object. However, let N be the sample number of the tested signal, p be the number of points on the log -log graph, obeying $N=2^p$, then one may observe that:

- p drastically diminishes when N decreases
- 1024 samples - i. e. for example 64 ms if Fs = 16 kHz - supply only 10 points; therefore these data characterise the shortest signal which may be studied by DyBC with Fs = 16 KHz.

Consequently, as it is impossible to highly increase Fs, DyCB is not suitable for the measurement of the dimension D of short signals. Thus, we designed a semi-continuous box counting method (SCBC) allowing to approach the small scales in case of short single dimension temporal signal. In this study, we describe this SCBC. Then, testing several mathematical signals, either with a known dimension, or which are not fractal, we compare its results in various cases to those obtained using the DyBC.

2 Methods and Material

2.1 Measurement methods

Let Ts be the sample duration i.e. 1/Fs. In this study, in order to first compare DyBC and SCBC, we used the lowest value allowing use of DyBC

- for Brownian signals: Fs=44.1 kHz, Sd=23 ms,N=1024
- for sinusoidal signal: Fs=16 kHz, Sd=64 ms, N=1024 i.e

Then, in order to appreciate the SCBC efficiency, we use various short duration signals (0.36, 0.72 and 1.45 ms for Brownian signals, and 1, 2 and 4 ms for sinusoidal signal) and compare the measured D with the theoretical D.

2.1.1 The dyadic box counting method

This method used 10 boxes whose sizes vary from 23 ms (1024 Ts) to 0.045 ms (2 Ts) for Brownian signals, and 64 ms (1024 Ts) to 0.125 ms (2 Ts) for sinusoidal signal.

We measured the slope on the log-log graph of these 10 points, which include all scales from 1024 Ts to 2 Ts (asD); we studied also the small scales from 8 Ts to 2 Ts, measuring only the slope of the 3 last points (ssD) on the log-log graph.

2.1.2 The semi-continuous box counting method

This method is directly derived from DyBC. It provides the first M points of the graph owing to the dyadic division of boxes at large scales up to a certain box size S_M, such as

$S_M = 2^M/Fs$. Then, at smaller scales, for each successive point the box size decreases by $1/Fs$, that provides the log-log graph with a large number of points. Thus, when $(N/S_{(M+x)}/Fs)$ is not a whole number, we have to peripherally and symmetrically reduce the analysed signal in the abscissa and the ordinate, so that a whole number is obtained. But these truncated samples are then reintroduced for designing the following boxes.

In the first part of this study, M=6, allowing to obtain a whole set of 36 points (Table I). We also separately measured asD, which is supplied by the slope of 36 points on the log-log graph , and ssD supplied by the slope of only the last 7 points, corresponding to the small scales from 8 Ts to 2 Ts.

Table I. Sizes of the 36 successive boxes of SCBC in case of D measurement of a 64 ms signal, owing to a 16 kHz frequency sampling, with M=6. For each size (i.e. duration) box, Bs-sn = box size in sample number.

In this example a dyadic division of box size is performed to obtain the first 6 points on the log-log graph, i.e. large scales, from 1024/Fs (64 ms) to 32/Fs (2 ms); then, for small scales, a 1/Fs (0.0625 ms) decreasing is realised from 31/Fs (1.9375 ms) to 2/Fs (0.1250 ms). However, when the ratio number of samples / duration box is not a whole number, the analyzed signal is peripherally and symmetrically reduced in abscissa and ordinate, so that a whole number is obtained. But these truncated samples are then reintroduced in the measurement of the following boxes. From the 36 boxes of this example, this table only indicates some data, and mainly: the values of the last 10 boxes; the values of the box corresponding to the biggest signal reduction of the signal (24 samples); in this case, for instance, for the measurement of this box (1.525 ms), the 12 first and 12 last samples of the 1024 samples which constitute the tested signal have not been analysed.

Dyadic division decreasing		1/Fs decreasing		Part of the signal which is not analysed with the box size Bs-sn
Box duration	Bs-sn	Box duration	Bs-sn	
64 ms	1028			
32 ms	512			
16 ms	256			
8 ms	128			
4 ms	64			
2 ms	32			
		1.9375 ms	31	1 sample
		1.8750 ms	30	4 samples
		1.8125 ms	29	9 samples
		1.5625 ms	25	24 samples
		0.6250 ms	10	4 samples
		0.5625 ms	9	7 samples
		0.5000 ms	8	0 sample
		0.4375 ms	7	2 samples
		0.3750 ms	6	4 samples
		0.3125 ms	5	4 samples
		0.2500 ms	4	0 sample
		0.1875 ms	3	1 sample
		0.1250 ms	2	0 sample
		0.0625 ms	1	0 sample

In the second part of our study, using various short duration signals, M=0 allowing to obtain 15, 31 and 63 points for 0.36, 0.72 and 1.45 ms in case of Brownian signal, and 1, 2, and 4 ms for sinusoidal signal. In these cases the slope of the log-log graph is based on the 10 last points.

2.2 Tested signals

We used 4 mathematical signals. Some of them are fractal with a known dimension: there are 3 Brownian signals (D = 1.5, 1.4, and 1.3). Another signal is not fractal (1 kHz sinusoidal signal), but its generalised dimension D_{gen} is known and equal to 1. For each signal we randomly selected 24 trajectories, each of 64 ms duration. We used Matlab software [6] to transform and analyse these sounds into 16 bit "wav" format.

Using both methods we measured each of these 24 D values and calculated the corresponding D mean value and confidence intervals.

2.3 Statistical study

Results were studied using the SPSS statistical package. Repeated measures of analysis of variance, using statistical contrasts to perform pairwise comparisons, were used to compare D measures.

Both DyBC and SCBC use the measure of the slope of the line obtained by linear regression on the log-log graph to determine D. However, only one of the two components is a variable, which represents the fluctuation in the signal. The other is a regular function of time. Therefore, in order to quantify the eventual bias, we calculated the slope and the corresponding error, i.e. the difference between the observed ordinate and the theoretical ordinate on the regression line.

3 Results

Measurement of slope and corresponding error gave values ranging from 10^{-4} to 10^{-7}. For each linear regression we observed a very small error ranged from 10^{-4} and 10^{-7}.

Table II. Mean value (upper line) and confidence interval (lower line) of D measurement of 24 randomly selected 64 ms duration parts of mathematic signals, using DyBC and SCBC, as a function of the size of the studied scales (all sizes = asD, from 23 to 0.045 ms for Brownian signals and 64 to 0.1250 ms for sinusoidal signal; only small sizes = ssD, from 0.18 to 0.045 ms for Brownian signal, and 0.5 to 0.125 ms for sinusoidal signal).

	DyBC - asD 10 boxes	SCBC - asD 36 boxes	DyBC - ssD 3 boxes	SCBC - ssD 7 boxes
Brownnian signal D=1.5	1.47 1.45 - 1.48	1.42 1.40 - 1.43	1.34 1.33 - 1.34	1.30 1.30 - 1.31
Brownnian signal D=1.4	1.39 1.37 - 1.41	1.36 ** 1.35 - 1.38	1.27 1.26 - 1.27	1.25 1.25 - 1.26
Brownnian signal D=1.3	1.30 1.29 - 1.33	1.28 1.27 - 1.30	1.21 1.20 - 1.21	1.25 1.25 - 1.26
1 kHz Sinus. signal D_{gen}=1	1.83 1.83 - 1.84	1.85 1.85 - 1.85	1.16 1.09 - 1.23	1.15 1.11 - 1.18

3.1 Fractal signals

Results are summarised in Table II.

In case of asD measurement, for Brownian signal, for which D = 1.5, the difference between DyBC and SCBC is significant. However none of these methods includes the theoretical D value in its confidence interval

For Brownian signals for which D = 1.4 and 1.3, the difference between DyBC and SCBC is not significant.

In case of ssD measurement, the difference between DyBC and SCBC is not significant for Brownian signal for which D = 1.5. It is significant for Brownian signal with D = 1.4 and 1.3. But the theoretical D value is never included in both methods' confidence intervals as the intervals do not overlap.

3.2 Non fractal signal

In cased of asD measurement, the difference between DyBC and SCBC is not significant. The calculated D is very different from D_{gen}.

In case of ssD measurement, the difference between DyBC and SCBC is not significant. But the calculated D tends to approach the D_{gen}.

Table III. Mean value (upper line) and confidence interval (lower line) of D measurement of 24 randomly selected parts of mathematical signals of various duration, using SCBC. NAS = total number of samples which have not been analysed.

	0.36 ms 16 Ts 15 points NAS=43	0.72 ms 32 Ts 31 points NAS=229	1.45 ms 64 Ts 63 points NAS=920
Brownnian signal **D=1.5**	1.31 1.27 - 1.34	1.33 1.31 - 1.35	1.30 1.28 - 1.32
Brownnian signal **D=1.4**	1.25 1.23 - 1.27	1.30 1.28 - 1.32	1.30 1.28 - 1.30
Brownnian signal **D=1.3**	1.24 1.23 - 1.26	1.28 1.27 - 1.29	**1.29** **1.28 - 1.31**

	1 ms 16 Ts 15 points NAS=43	2 ms 32 Ts 31 points NAS=229	4 ms 64 Ts 63 points NAS=920
1 kHz Sinus. signal **D_{gen}=1**	1.32 1.31 - 1.33	1.19 1.18 - 1.20	1.15 1.14 - 1.15

3.3 Short duration signals

The measured D value progressively approaches the theoretical D as the duration of the signal increases (Table III). In case of Brownian signals with D = 1.4 and 1.3, the differences observed between each pair of measures are significant only for the comparisons 0.36 ms vs 0.72 ms and 0.36 ms vs 1.45 ms. In case of non fractal signal the difference is significant for the comparisons 1 ms vs 2 ms, 1 ms vs 4 ms, 2 ms vs 4 ms.

4 Discussion

4.1 Methodology

We could have tried to study other fractal signals than Brownian signals. However, in practice it is difficult to obtain such mathematical signals, which are surely fractal; this difficulty is evident if, for instance, we consider white noise.

The value of M is important and we shall discuss it later.

4.2 Efficiency of SCBC

4.2.1 Fractal signals

In case of asD measurement, one may observe that the efficiency of both DyBC and SCBC decreases as D - i.e. the signal roughness - increases. That implies that the signal irregularities are not correctly taken into account by the relatively narrow window which constitutes the tested signal duration.

Besides, in case of ssD measurement, despite the difference between SCBC and DyBC being significant, none of these methods gives correct D value. That is probably due to the fact that, in this particular case of fractal signal, the observation of only the small scales does not include the long dependence of the Brownian signal.

4.2.2 Non fractal signal

The values of asD measurement are very different from the true value of D_{gen}. That seems normal, because the concept of D_{gen} is valuable only on infinitely small scales.

Conversely, ssD tends to reach the true value of D_{gen}, and we may observe that there is no significant difference between DyBC and SCBC

4.2.3 Short duration signal

If we consider Table III and the SCBC-asD column of Table II, and recall the significance of the difference between the various pairs of measurements, we may observe that the efficiency of SCBC depends on the signal duration, whatever the signal. In case of Brownian signal, this efficiency also inversely depends on the D value. These discrepancies may be explained by:

- the truncation of the Brownian signal, which distorts its correlation function adding high frequencies in its spectrum,
- the temporary suppression of some samples during the SCBC management. This suppression represents a bias; this bias probably increases with the number of samples which are not analysed, and also with the reduction in signal duration. Here the value of M must be discussed. In this preliminary study we arbitrarily chose M=6 for comparison between SCBC and DyBC, and M=0 for testing efficiency of SCBC in case of signals shorter than 64 ms. However, the fewer are the samples which are not analysed, the better is probably the SCBC efficiency. It could be worthwhile to improve SCBC efficiency choosing M value as great as possible, in such a way as to obtaining only 10 necessary points for slope calculation.

Nevertheless one must underline that, in case of non fractal signal, the value of D_{gen} is correctly approached by SCBC.

4.2.4   ssD significance

One may observe that the use of asD is efficient for signals with long correlations (e.g. Brownian signals), while ssD is more efficient for signals where the dimension is observable at the smallest scales (e.g sinusoidal signals) or for signals that are uncertain to have a fractal dimension. But fractal dimension is always defined as a limit, which implies that the structure can be analysed in arbitrarily high resolution. However, this trend is only achieved for an infinite number of points, which is conceivable for a theoretical signal, but impossible for a natural signal. In practice, if we consider a physical object, its structure is given by a finite data set representing its discrete digitisation. Thus, using only the last points of the curve, i.e the smallest scales, the dimension of an object which is fractal is not measured exactly. But ssD may be considered as a kind of signature of the signal, which may be characteristic [4] in some cases; the advantage of this estimation of ssD is that it highlights the convergence to generalised dimension of an object for the smallest resolutions. That explains why, in case of sinusoidal signal, we observed that the measurement of ssD gives the trend towards the correct D value more rapidly (Fig.1,2,3,4).

Figure 1. D values (ordinate) obtained using box-counting method for several sinusoidal sounds as a function of frequency sampling using asD measurement.

D
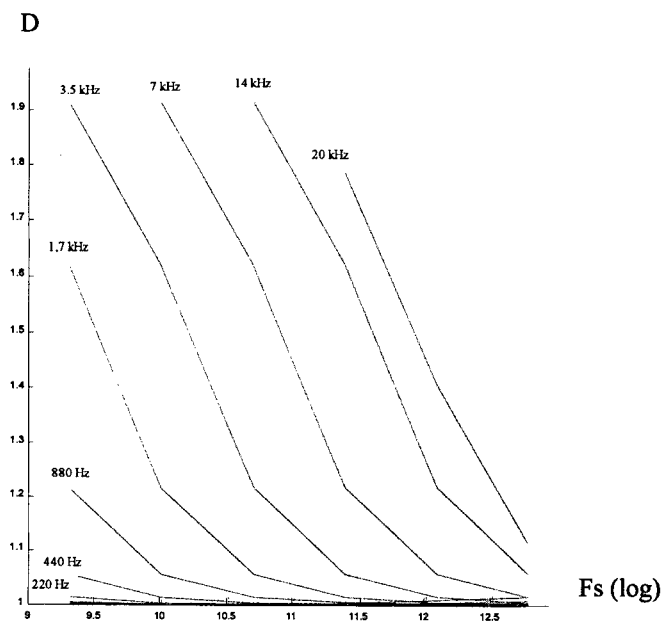
Figure 2. D values (ordinate) obtained using box-counting method for several sinusoidal sounds as a function of frequency sampling using ssD measurement.

D

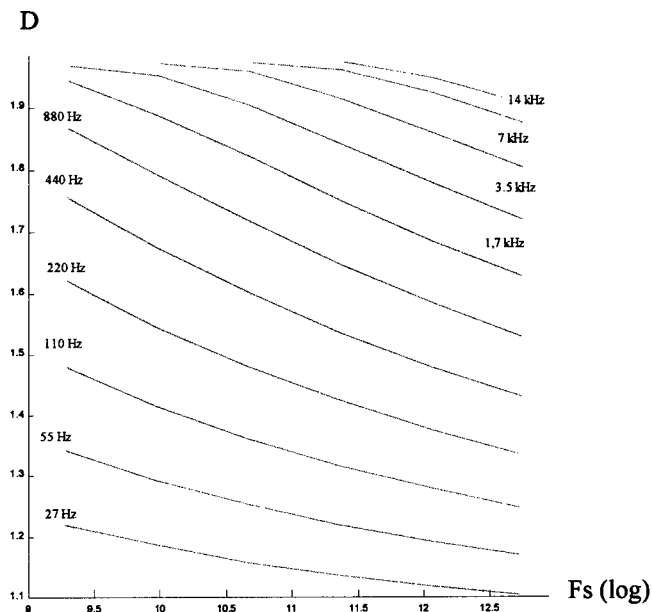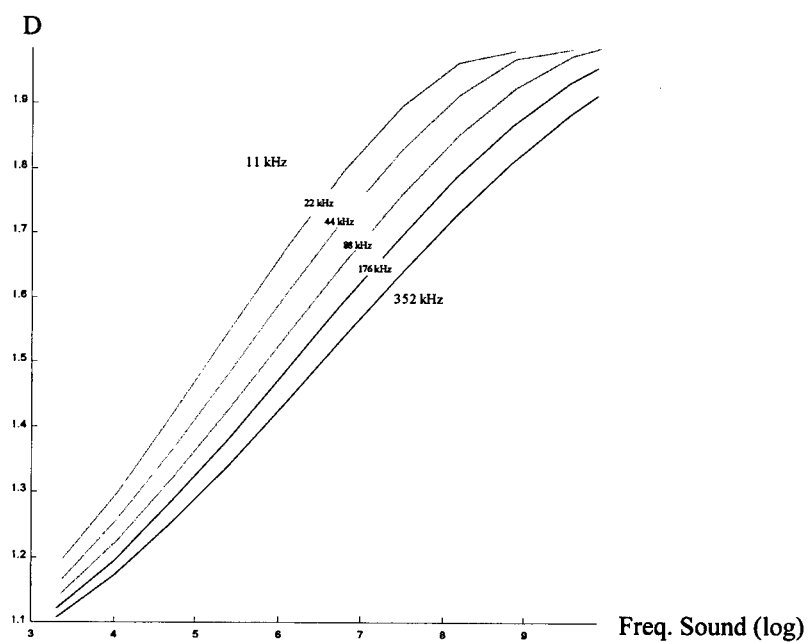**Figure 3.** D values (ordinate) obtained using box-counting method for several sinusoidal sounds as a function of frequency sound using asD measurement.

D
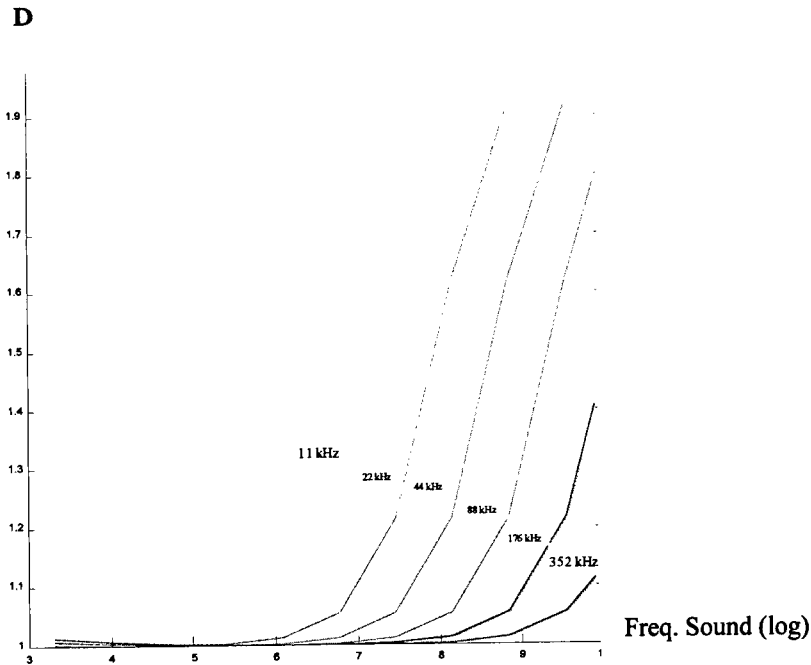
Figure 4. D values (ordinate) obtained using box-counting method for several sinusoidal sounds as a function of frequency sound using ssD measurement.

4.3 In the literature

One may find other methods in the literature to determine D of a single dimension temporal signal. In our Laboratory, we attempted to employ these methods [7] using the same mathematical signals as we used in this paper. All these results may be compared.

4.3.1 The Richardson length method [8]

This method measures the perimeter of an object with various length of rulers; when plotting log perimeter against log ruler length, a fractal object gives a straight line with a negative slope S, and D = 1 - S. But this method is difficult to be computed without any risk of bias. Moreover most signals must be digitised to be easily studied using computers; consequently they are not yet continuous, burepresent a series of data. Anyway Pickover [9] demonstrated that this method supplies with the same results as using box-counting method on the same but digitised signal.

4.3.2 The power spectrum density (PSD) measurement

This technique has been employed [10] using the Fourier transform generated spectral density function, which gives intensity or power at each frequency. The results demonstrate that this method is valuable for sounds with PSD in 1/f, as Brownian signals for instance. But it is not usable for signals analysis, which have not PSD in 1/f.

4.3.3   The rescaled range analysis (RRA)

This is the earliest empirical method. It has been described by Hurst to study the long time series of Nile floods. It is based on the assumption that a large number of natural phenomena are time series with a long-term correlation. This method was recently studied by Bassingthwaighte *et al.* [11]. It allows to characterise a one-dimensional time series by simultaneously providing a measure of variance and long-term correlation of its components (the term « memory » is often used). But the results obtained with the signals where the PSD is not a power law, or which do not have - even theoretically - a fractal dimension are totally incoherent. This is the case for a constant or a sinusoidal signal. RRA only appeared to be effective for the signals with a very long correlation, or a power spectrum of 1/f. However, even in these cases, application of known Brownian signals underestimated true H for H>0.72, and underestimated H for H<0.72.

4.3.4   The dispersional analysis method (DAM)

This method is also empirical. It was recently studied by Bassingthwaighte *et al.* [12] to evaluate whether it was effective to determine the fractal dimension of a Brownian signal, according to the type of this signal and the size of the data set. DAM measures the standard deviation of the single dimension signal at different scales. The objective is to reveal a possible power law in the successive values of the standard deviation of the signal. But for the signals where the PSD is not a power law, or which do not have a fractal dimension, the results are similar to those obtained with RRA. Consequently, despite the fact that the evaluation of true H of Brownian signals is more precise than with RRA, one may conclude that all these methods are less efficient than box DyBC or SCBC.

4.3.5   Other tools

Although the aim of this paper is not the study of the eventual fractal features of speech, one may briefly mention other tools proposed by several authors for use of fractal geometry as a model for describing irregularities of graphical wave forms of human speech. Bohez *et al.* [13] used box counting method and cluster analysis; however Ds values have not been clearly specified. Using different box shapes and a specific box-counting algorithm, Maragos [14] described different D (1 to 1.3) for vowels as a function of their scale analysis, and his results are in accordance with our first measurement [4] . Other authors studied the multifractal and chaotic features of speech, but we shall not discuss them in this paper, which only considers a fractal approach to single dimension, short temporal signals.

5   Conclusion

As SCBC increases the number of points at small scales, it improves the DyBC performances, pushing far away its physical limits due to frequency sampling and duration of the tested signal. However its efficiency presents comparable limits:
- some of them are physical, and for example Fs=16kHz, at least 10 points on the log-log graph are necessary, the shortest signal which may be analysed using SCBC must have a minimum of 0.6875 ms duration.
- others are due to the principle of the method, which leads to some samples not being analysed. However, choosing M value as a function of Sd reduces the number

of points necessary for the slope calculation to 10 on the log-log graph could improve SCBC efficiency.
We plan to study further this possible improvement. Then we shall use this method to explore the eventual dimension D of transient parts of consonants.

6 Acknowledgements

References

1. Chouard CH, Ouayoun M, Meyer B, Fugain C. Speech coding strategies of the Digisonic fully digitized cochlear implant. Acta Otolaryngol (Stockh) 1995; 115: 264-268.
2. Chouard CH, Ouayoun M, Meyer B, Secqueville T, Bachelot G, Génin J. Auditory performances of a 3-4-7 programmable numeric filters prosthesis. Audiology 1996; 35: 23-36
3. Ouayoun M, Pean V, Genin J, Bachelot G, Fugain C, Meyer B, Chouard CH. Asynchronous Interleaved Stimulation (AIS): a new speech coding strategy for cochlear implant. *Acta Otolaryngol (Stockh)* 1997; 117: 182-186.
4. Ouayoun M, Péan V, Meyer B, Chouard CH. A study of speech fractal dimension.. *Acta Otolaryngol (Stockh)* 1999; 119: 261-266.
5. Robinson A. Fractal fingers in viscous fluids. *Sciences* 1985; 228: 1080-1085.
6. MATLAB 5.2 – The MathWorks Inc. copyright 1984 – 1998.
7. Reznik L, Dubos D. Etude de la fractalité de la parole. Mémoire d'ingénieur de 2ième année réalisé au Laboratoire de Recherche ORL du CHU St Antoine. 1998. Ecole Française d'Electronique et d'Informatique. 30 Avenue de la République. F-94800, Villejuif
8. Smith TG, Lange GD, Marks WB. Fractal methods and results in cellular morphology--dimensions, lacunarity and multifractals. *J Neurosci Methods* 1996 Nov: 69(2): 123-136.
9. Pickover CA, Khorasani A. Fractal characterization of speech waveform graphs. *Comput & Graphics* 1986; 10: 51-61.
10. Clarke J, Voss R. "1/f" noise in music and speech. Nature 1975; 258: 317-318.
11. Bassingthwaighte JB, Raymond GM. Evaluating rescaled range analysis for time series. *Annals of Biomedical Enginering* 1994; 22: 432-444.
12. Bassingthwaighte JB, Raymond GM. Evaluation of the dispersional analysis method for fractal time series. *Annals of Biomedical Enginering* 1995; 23: 491-505.
13. Bohez EL, Senevirathne TR, Van Winden JA. Fractal dimension and iterated function system for speech recognition. Electronics letters 1992; 28: 1382-1384.
14. Maragos P. Fractal aspects of speech signals: dimension and interpolation. *Proc IEEE IACSSP* Toronto 1991; 1: 417-422.

A FRACTIONAL BROWNIAN MOTION MODEL OF CRACKING

P.S. ADDISON, L.T. DOUGAN, A.S. NDUMU, W. M.C. MACKENZIE

Civil Engineering Group, School of the Built Environment, Napier University, Merchiston Campus, 10 Colinton Road, Edinburgh, EH10 5DT, Scotland, UK.

An attempt is made to find the fractal cutoff of crack profiles on the tension face of concrete beams subjected to uni-axial bending. Previous work by the authors has shown that such cracking can be interpreted as a non-Fickian diffusive phenomenon resulting from a self-affine random fractal process: specifically fractional Brownian motion (fBm). In addition, a spatial description of the cracking geometry can be found from experimental data using both a (Hurst) scaling exponent and a diffusion-type coefficient. Herein the authors find that the fractal description of the crack profiles extends down to less than 0.75μm. The use of a scanning electron microscope to probe the crack profile (and surface) at smaller scales is discussed and the synthesis of crack surfaces using fBm is described briefly.

1 Introduction

An understanding of the behaviour of cracking in structural elements is of great importance in the analysis and subsequent safe design of engineering structures. As yet, however, there is no definitive theoretical framework for the propagation of cracks and resulting fracture energy. It has recently been found that the irregular geometries of both crack surfaces and crack profiles in a variety of materials may be described (and subsequently modelled) using fractal geometry. The use of fractal geometry to describe cracking phenomena is now widespread (e.g. see references[1-4]). Previous work by the authors[5,6] has shown that crack profiles on concrete beams in tension can be modelled as fBm trace functions which require a Hurst exponent and a spatial diffusion coefficient to completely describe the spatial distribution of the crack. In addition, the authors have linked the fBm description of the cracking phenomena to an effective Fokker-Planck equation which described the diffusive nature of the cracking phenomena through space. In this paper the cracks are investigated at higher resolutions in an attempt to determine whether a fractal cut-off scale exists. The value of such a Euclidean threshold is of significant importance in the determination of the energy of fracture.

2 The Diffusive Nature of fBm

Fractional Brownian Motion (fBm) is a generalisation of Brownian motion suggested by Mandelbrot[7] which has found a variety of uses in the natural sciences (see for example Addison and Ndumu[8] and the references contained therein). Fractional Brownian motion is defined as:

$$y(x) = \frac{1}{\Gamma\left(H+\frac{1}{2}\right)}\left\{\int_{-\infty}^{0}\left[(x-x')^{H-\frac{1}{2}}-(-x')^{H-\frac{1}{2}}\right]dW(x')+\int_{0}^{x}(x-x')^{H-\frac{1}{2}}\,dW(x')\right\}$$

(1)

where $dW(x)$ is a Gaussian random function with zero mean and unit standard deviation, H is the Hurst exponent[9], and Γ is the gamma function. When $H=0.5$ Eq. (1) models classical Brownian motion which produces normal, or Fickian, diffusion. From Eq. (1) it may be seen that the fBm process is correlated over all length scales, i.e., it has an infinite memory associated with it.
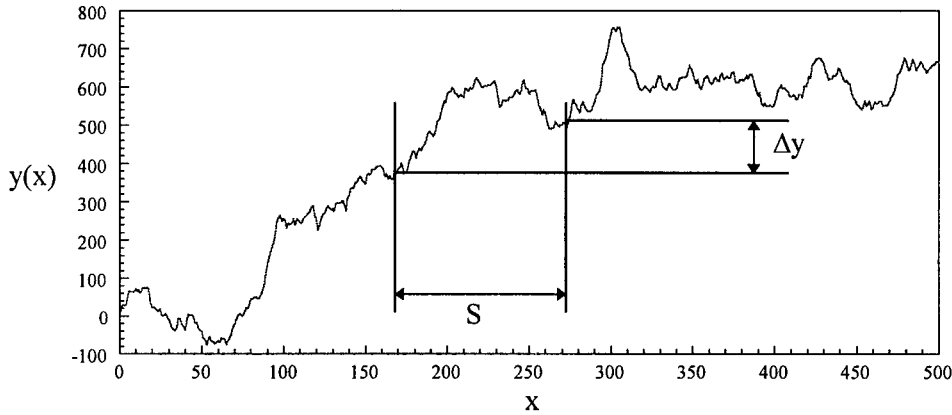
Figure 1: Diffusive scaling of an fBm ($H=0.75$ $K_f=10$)

An example of a superdiffusive fBm (i.e. one with $H > 0.5$) is shown in Fig. 1. The diffusive scaling of the fBm process shown in Fig. 1 may be defined as

$$\sigma_y = \sqrt{2K_f}\ s^H \qquad (2)$$

where σ_y is the standard deviation of the y-excursions (Δy) on the trace for a window length s; K_f is a fractal diffusion coefficient. Eq. (2) is in fact the standard deviation of the probability density function

$$P(y,x) = \frac{1}{\sqrt{4\pi K_f x^{2H}}} \exp\left(-\frac{y^2}{4K_f x^{2H}}\right) \qquad (3)$$

which is a non-Fickian scaling of a Gaussian probability density function through space. (If $H = \frac{1}{2}$ then Eq. (3) reduces to the solution of a Fickian based diffusion from a point source.) Furthermore, it has been shown by Wang and Lung[10] that Eq. (3) is the solution to the effective Fokker-Planck equation:

$$\frac{\partial P(y,x)}{\partial x} = 2H\ K_f\ x^{2H-1} \frac{\partial^2 P(y,x)}{\partial y^2} \qquad (4)$$

which describes the probability of occurrence of $y(x)$ at spatial location x. Eq. (4) is in fact a generalisation of the classical Fickian diffusion equation with a spatial diffusion coefficient. The equation reduces to the classical equation for $H=0.5$.

It can be seen from the above that, over a large number of realisations, fBm approximates a non-Fickian diffusive process described by Eq. (4). The authors have previously shown that persistent fBm ($H > 0.5$) is a suitable model for cracking on the tension face of a concrete beam in bending[5]. In addition, both H and K_f are required for a complete geometric description of the cracking phenomena. It was shown by the authors how these parameters can be found from experiment. The mean values of K_f and H were found for a series of flexure cracks in concrete beams to be 0.084 and 0.77, respectively (i.e. superdiffusive surfaces with fractal dimensions between 1 and 1.5). This gives the standard deviation of the cracking across the beam as $\sigma_y = \sqrt{2 \times 0.084}\ s^{0.77}$ where both s and σ_y are expressed in millimetres. Thus, for the 40mm wide specimens used in the study the expected standard deviation of the crack displacement across the beam is 7.02mm. These experimentally derived parameters can be used to synthesise crack patterns using fBms. An example of this is shown in Fig. 2 using the fBm generation method described by Addison et al.,[11]. In the figure a crack with measured values of H and K_f of 0.75 and 0.133 respectively is shown together with a synthesised crack with the same parameter values. The similarity between the two traces is evident from a visual inspection of the plot.
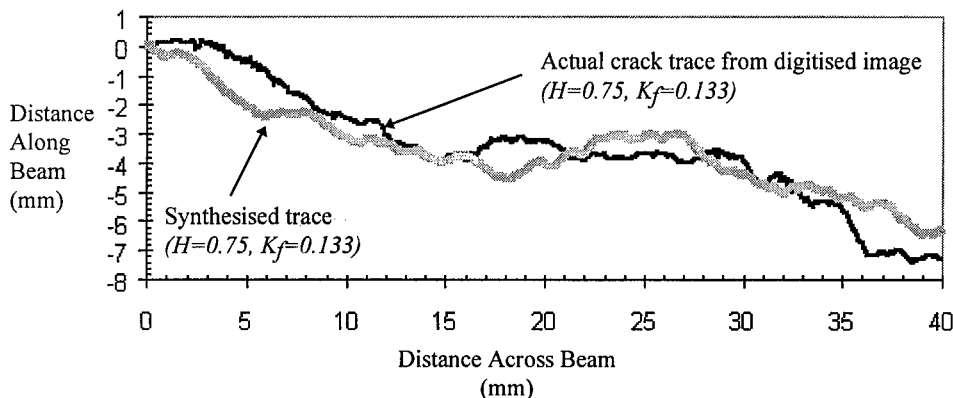
Figure 2: Comparison of synthesised and experimental cracks

3   Crack Profile Analysis

Natural fractals tend to exhibit fractal characteristics over a limited range of scales[12]. Below a cut-off level, the natural fractal object tends to revert to a Euclidean form. The authors have recently pursued the search for the cut-off length scale in the crack patterns as it has implications for the measurement of the true areas of crack surfaces and hence energy dissipation across the surface. Fig. 3 shows one of the cracks studied by the authors at a magnification level of 6×. Seven boxes are placed on the crack profile indicating regions where a closer inspection was taken of the crack profile at the higher magnification of 40×. In addition, two smaller boxes indicate locations where the crack was studied at 50× and 100× magnification, respectively. Fig. 4 contains a log-log plot of σ_y against s. From such a plot it is possible to calculate both H and K_f. A line of slope $H=1$ is also given in the plot, corresponding to a dimension of unity, i.e. a smooth Euclidean curve. It would be expected that the plotted curves tend to this slope at the Euclidean cut-off.
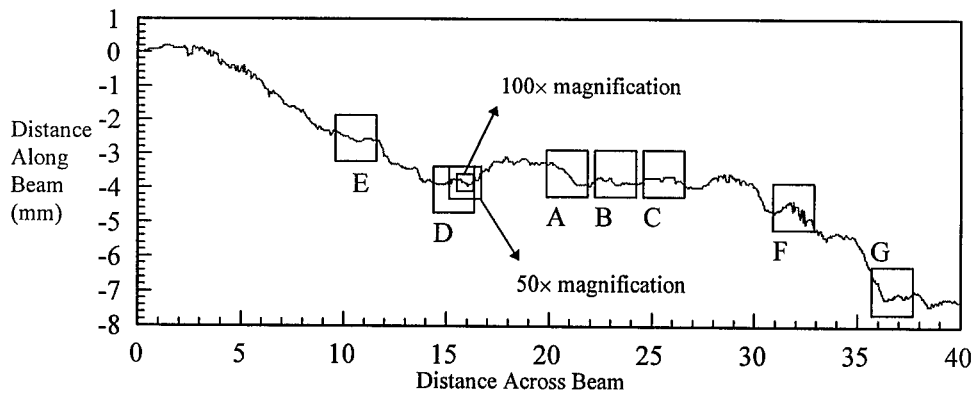
Figure 3: Crack profile showing analysed regions
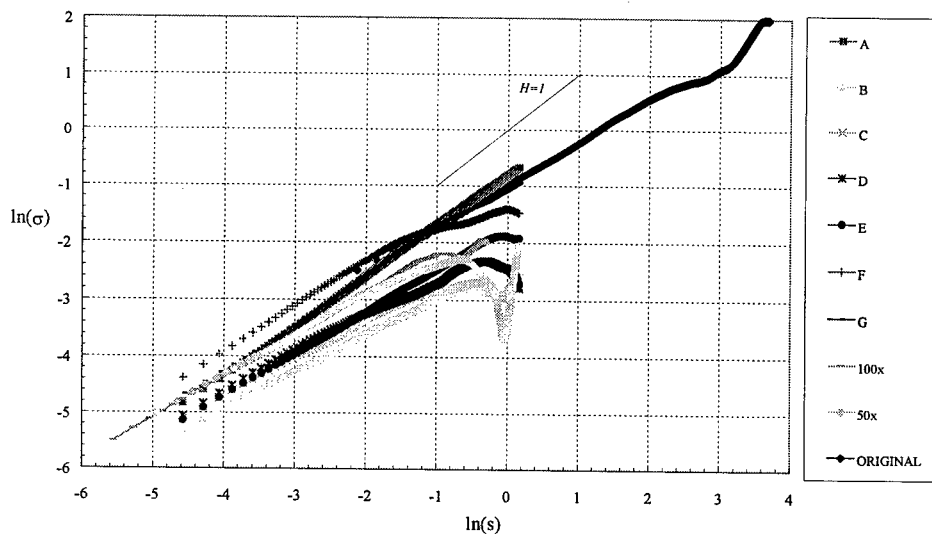
Figure 4: Logarithmic plots of σ_s versus s for the crack shown in figure 3. The graph shows (extending out, bottom left) two sections representing 50× and 100× magnifications.

Table 1 contains the H and K_f values for the various regions of the crack in Fig. 3 measured from Fig. 4. It can be seen from the table that measured H values for the whole crack and the average values from the seven boxes A to G are in good agreement. However, the K_f values between the two are significantly different. In fact the K_f values vary over a large range (0.004-0.118) across the selected boxes. The reason for this variability is as yet unclear. Zooming in at 50× and 100× again produces persistent values of the Hurst exponent. It can be seen from the plot that the fractal description of the curves extends down to the lower limits of the 100× magnification: this relates to a resolution of 7.5×10^{-4} mm of crack per pixel.

Table1: The Measured H and K_f Values for the Crack in Figure 3.

	H	K_f
Whole Crack (6×)	**0.74**	**0.069**
A (40×)	0.87	0.093
B (40×)	0.64	0.004
C (40×)	0.68	0.008
D (40×)	0.61	0.018
E (40×)	0.72	0.015
F (40×)	0.82	0.118
G (40×)	0.83	0.091
Average (A-G)	**0.74**	**0.050**
50×	**0.71**	**0.025**
100×	**0.74**	**0.030**

4   Concluding Remarks

From Fig. 4, the lower limit to the fractal behaviour of the cracking patterns (if it exists!) appears to be below 0.75μm. This is significantly less than the value between 10 and 20 μm (which is the approximate size of calcium silicate hydrate) suggested by Souma and Barton[4]. The authors have recently initiated research to search for a fractal cut-off scale at higher resolutions at the crack edge using a scanning electron microscope (e.g. Fig. 5). This work has so far proved inconclusive due to the difficulty in finding reasonable vertical sections through the crack edge at these higher scales. It is hoped that an improvement in the experimental techniques will lead to a better understanding of the crack geometry at these smaller scales.
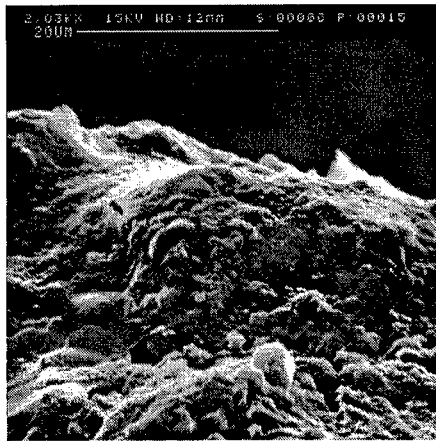
Figure 5: An electron microscope image of the crack profile at 2000× magnification. Note the difficulty in defining the edge of the crack.

122

The determination of the true fractal cut-off scale has implications for the synthesis of prefractal fBm crack profiles and surfaces and hence the calculation of crack energy across the crack surface. If the crack surface is in fact fBm, then the crack profile may be treated as the result of a vertical cut through the surface[13]. If this proves to be the case, then it is known that the fractal dimension of the surface $D_{surface}$ is equal to $D_{profile}+1$. It is relatively simple to generate such a surface using a variety of methods. It should be possible therefore to synthesise the crack surface using the K_f and H values found from experiment. Fig. 6 shows an fBm surface generated using the turning bands method[14]. The authors intend to pursue the measurement and synthesis of crack profiles and surfaces in order to define the fracture energy of cracking in terms of a fractal geometric framework based on fBms.
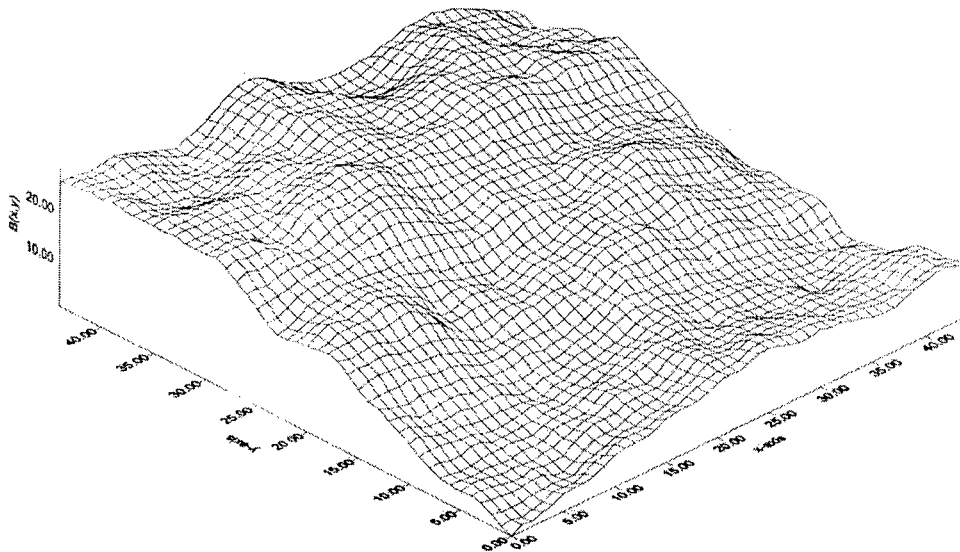
Figure 6: An fBm surface generated using the turning bands method ($H=0.8$)

References

1. Carpinteri A. (1994a). 'Scaling laws and renormalization groups for strength and toughness of disordered materials.' *Int. J. Solids Struct.*, 31(3), 291-302.
2. Carpinteri A. (1994b). 'Fractal nature of material microstructure and size effects on apparent mechanical properties.' *Mechanics of Materials*, 18, 89-101.
3. Borodich F.M. (1997). 'Some fractal models of fracture.' *J. Mech. Phys. Solids*, 45(2), 239-259.
4. Saouma V.E., Barton C.C. and Gamaleldin N.A. (1990). 'Fractal characterisation of fracture surfaces in concrete.' *Engng. Fract. Mech.*, 35, 47-53.

5. Addison P.S., MacKenzie W.M.C., Ndumu A.S., Dougan L. and Hunter R. (1999), 'Fractal Cracking of Concrete: Parameterisation of Spatial Diffusion', *ASCE J. Engng. Mech.*, 125(6), 622-629.

6. Addison P.S., Dougan L.T., Ndumu A.S and Mackenzie W.M.C. (1999). 'A Complete Geometric Description of Cracked Concrete', 13th ASCE Engineering Mechanics Division Conference, Baltimore, MD, USA, June 13-16.

7. Mandelbrot B.B. & Wallis J.R. (1969). 'Computer experiments with fractional Gaussian noises. Part 3, Mathematical Appendix.' *Water Resour. Res.*, 5, 260-267.

8. Addison P.S. and Ndumu A.S. (1999), 'Engineering Applications of Fractional Brownian Motion: Self-Similar and Self-Affine Random Processes', *Fractals*, 7(2), 151-157.

9. Hurst H.E. (1951). 'Long term storage capacity of reservoirs', *Trans. Am. Soc. Civil Eng.*, 116, 770-808.

10. Wang K.G., and Lung, C.W. (1990). 'Long-time correlation effects and fractal Brownian motion.' *Phys. Letts. A*, 151(3,4), 119-121.

11. Addison P.S., Qu B, Ndumu A.S. and Pyrah I.C. (1998). 'A particle tracking model for non-Fickian subsurface diffusion.' *Math. Geology*, 30(6), 695-716.

12. Bunde A & Havlin S., eds. (1994). Fractals in science. Springer-Verlag, Berlin.

13. Addison P.S. (1997). Fractals and chaos: An illustrated course. Institute of Physics Publishing, Bristol.

14. Mantoglou, A., and Wilson, J. L. (1982). The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resour. Res.*, 18(5), 1379-1394.

SELF-AFFINE SCALING STUDIES ON FRACTOGRAPHY

J. M. LI, L. LU AND M. O. LAI

Department of Mechanical and Production Engineering
National University of Singapore
10 Kent Ridge Crescent, Singapore 119260
E-mail: mpeluli@leonis.nus.edu.sg

Applying variation-correlation method to images of fractography obtained from the scanning electron microscope (SEM), it has been found that there exists a fractal characteristic length within which the fractured surfaces are fractal. Investigation shows that the fractal characteristic length can represent the statistical maximum size of texture of the SEM image. Multi-magnification fractal analysis has shown that fractography cannot be described by a single fractal dimension but rather a series of fractal dimensions. Fractal study on the fractured surfaces of gray iron with different grain size has shown that there exists a positive relationship between tensile strength and fractal dimension. However, no essential relationship between impact toughness and fractal dimension for HP26Cr35Ni alloy could be obtained. Fractal dimension is sensitive to changes of mechanical properties caused by geometric factors such as grain size and is suitable to quantitatively describe the irregularity of fractography. For a given fracture mode, fractal dimension is not universal.

1   Introduction

Fracture phenomenon is one of the most intriguing areas in material science and engineering because it is associated with the life and safety of engineering parts. There is a lot of information related to fracture mechanism in fractography. Fractal based studies on fractography become important because of the problem of variance of conventional quantitative parameters with scales. During the last twenty years, many fractal-based methods for quantitative analysis on fractography have been developed. The methods can be categorized into the following three classes: (a) slit island analysis [1]; (b) profile analysis [1-5]; and (c) 3-dimensional (3D) surface analysis [6-8]. Among them, 3D surface analysis is the most interesting. The fractal characteristics of fracture surfaces and the relationship between fractal parameters and mechanical properties have been investigated using these methods. Some results justify that fractal dimension can be a measure of the toughness of a material [1, 9, 10]. On the other hand, some results have shown the existence of a universal value of fractal dimension for a given fracture mode. It has been conjectured that the value of fractal parameter α ($\alpha=3-D$ where D is fractal dimension) is 0.8 for ductile fracture [11] and 0.87 for brittle fracture [3]. These values imply that there is no correlation between fractal dimension and toughness because a universal value denies any dependence of fracture features on a specific toughness. In this paper, a new 3D surface model based on variation profile analysis has been introduced. Using this model, the fractal characteristics of different fracture modes and the correlation between fractal dimension and mechanical properties have been studied.

2   Experimental Procedures

A typical ductile fracture of microvoid coalescence was obtained by tensile testing a specimen of Assab 760 medium carbon steel. A transgranular cleavage mode was obtained by fracturing a Charpy V-notch specimen of the same steel after immersing it in

the liquid-nitrogen for 60 minutes. Fracture of a pure Al_2O_3 ceramic specimen appeared to be brittle intergranular mode. To analyze the influence of geometric and non-geometric factors on fractal dimension of fractography, specimen A of gray iron was modified with 0.4%wt 75SiFe inoculation, while specimen B, with 0.3%wt FeSiCuMo metal stream inoculation to improve the graphite-eutectic microstructures. Tensile strength σ_b was obtained by tensile testing the various materials. A series of specimens of HP26Cr35Ni alloy with the size 4x4x35(mm) were taken from same regions of solidification. Thermal shocks before mechanical testing were carried out 0, 5, 10, 15 and 20 times. In the thermal shock, the specimens were heated to 1100°C for 10 minutes followed by water-quenching. After the thermal shock, impact toughness was obtained by impact testing without Charpy V-notch. The nominal compositions of these materials are listed in Table 1. All fracture surfaces were imaged by SEM followed by digitizing.

Table 1. Composition of materials investigated in wt%.

Materials	C	Si	Mn	P	S	Cr	Ni	Fe
760 steel	0.5	0.25	0.5					Bal.
Gray iron	3.0	0.12	0.7	<0.3	<0.05			Bal.
HP26Cr35Ni	0.55	2.5	2.0	<0.04	<0.04	26	35	Bal.

3 Variation-Correlation Method

The variation method has been introduced by Dubuc *et al* [4]. In this method, an ε-oscillation for the height function $h(x)$ of a profile, $v(x, \varepsilon)$, is defined as follows:

$$v(x,\varepsilon) = h_{\max}(x) - h_{\min}(x), \qquad |x - x_0| < \varepsilon, \tag{1}$$

The integral $V(\varepsilon)$ may be defined as:

$$V(\varepsilon) = \int v(x,\varepsilon)dx. \tag{2}$$

$V(\varepsilon)$ is termed the ε-variation. The limit of $V(\varepsilon)$ is zero for $\varepsilon \to 0$ due to continuity of the function $h(x)$. The rate at which $V(\varepsilon)$ scales when $\varepsilon \to 0$ is:

$$V(\varepsilon) \propto \varepsilon^{2-D}, \tag{3}$$

which provides a method to calculate the fractal dimension D. The variation method can be interpreted as a numerical technique of box counting [4]. It has been demonstrated to be more accurate than any other techniques for fractal analysis of self-affine curves [4,5].

Based on variation method, a variation-correlation model for 3D surface has been developed [8]. In this model, it is assumed that $H(x,y)$ is a dimensionless field-like variable of a surface which may denote distribution of height or angle. If $H_{max}(x,y,\varepsilon)$ and $H_{min}(x,y,\varepsilon)$ are the local maximum and minimum of $H(x,y)$ in an ε-neighborhood respectively, a variation-correlation function $V_{cor}(\varepsilon)$ is defined as follows

$$V_{cor}(\varepsilon) = <[H_{\max}(x,y,\varepsilon) - H_{\min}(x,y,\varepsilon)]^2> \tag{4}$$

The symbol <...> refers to two consecutive averaging operations. If the variation-correlation function $V_{cor}(\varepsilon)$ of the variable $H(x,y)$ on a surface obeys a power law in some significant range of scales $\varepsilon_{min} \leq \varepsilon \leq \varepsilon_{max}$, i.e.,

$$V_{cor}(\varepsilon) \propto \varepsilon^{2\alpha} \tag{5}$$

where the fractal parameter $\alpha = 3-D_{cor}$, the variation-correlation is called fractal variation-correlation function.

4 Results and Discussion

4.1 Fractal Characterization of Different Fracture Modes

The typical fracture modes of microvoid coalescence, transgranular cleavage and brittle intergranular fracture are shown in Fig.1 (a), (b) and (c). Fig.1 (d), which gives the V_{cor}-ε curves of these fracture surfaces, shows that the variation-correlation appears as lateral growth only within the range $\varepsilon < \varepsilon_{max}$. When $\varepsilon > \varepsilon_{max}$, the V_{cor}-ε curves appear to reach plateau value. These observations imply that the following short- and long-range relationships of variation-correlation versus scale are true

$$V_{cor}(\varepsilon) \propto \begin{cases} \varepsilon^{2\alpha} & \varepsilon \ll \varepsilon_{max} \\ V_{max}^2 & \varepsilon \gg \varepsilon_{max} \end{cases} \tag{6}$$

From Eq.(6), the log-log plot of V_{cor} versus ε should be a straight line in the short-range of scales, of which the slope is different from that in the long-range of scales. To estimate ε_{max}, a linear measure M_F is introduced as follows

$$M_F = \frac{1-\delta}{1+\delta} \tag{7}$$

where δ is the degree of disperse expressing the fluctuation of the set of points (x_i, y_i) from the fit line. δ is defined as

$$\delta = \sqrt{\frac{\beta - \rho^2}{\beta}}$$

where

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad \beta = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

(a) (b) (c)
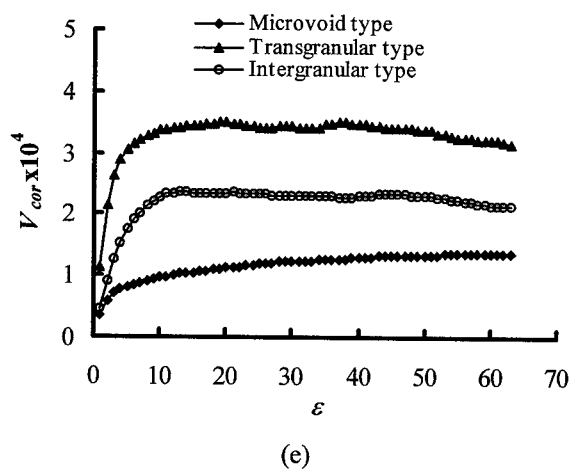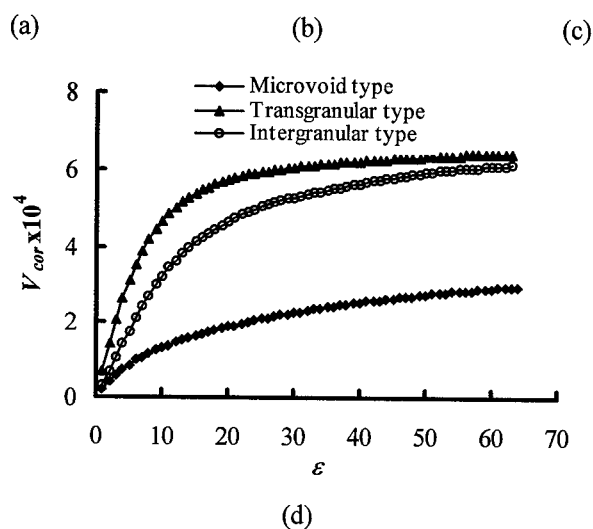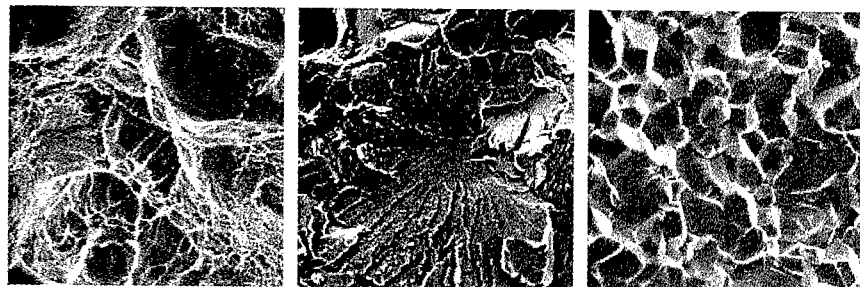

(d)

(e)

Fig.1. SEM fractography and fractal characteristics. (a) Microvoid fracture; (b) Transgranular cleavage fracture; (c) Brittle intergranular fracture; (d) Variation-correlation versus scale curve based on variation-correlation analysis; (e) Height-correlation versus scale curve base on height-height correlation analysis. The resolution of the images is $68.75\mu m/200$pixels.

The value of M_F falls in the range between 0 and 1. When M_F tends to 1, the distribution of the points (x_i, y_i) is perfectly linear. When M_F tends to 0, the distribution becomes non-linear. Table 2 shows the values of D and ε_{max} estimated with M_F=0.7, while the correlation coefficient r of log-log plot is over 0.98, indicating M_F is a better linear measure than r.

Table 2. The estimation of fractal parameters and the maximum size d_{max} of the basic micro-morphology element on the fractography

Fractography	Variation-correlation				Height-height correlation				D_{max}
	D_{cor}	ε_{max}	r	M_F	D_{h-h}	ε_{max}	r	M_F	
Microvoid type	2.58	64	0.984	0.7	2.68	5	0.982	0.7	75
Transgranular type	2.54	17	0.983	0.7	2.58	5	0.972	0.7	31
Intergranular type	2.44	26	0.983	0.7	2.33	10	0.982	0.7	35

The above results confirm that there exists a characteristic length ε_{max} so that the log-log plot of variation-correlation is linear within ε_{max} for the fractography of a real material. Different from an ideal self-affine profile, for example, Weierstrass-Mandelbrot curve whose V_{cor} appears as lateral growth at all possible scales, the morphologies of fracture of real materials are constituted by those basic micro-morphology elements such as dimples, cleavage facets or boundary facets. The basic micro-morphology elements are affine at a limited scale, to form the texture of SEM images. Assuming that V_{max} is the expected value of maximum variation of the texture at all possible scales, V_{max} is bounded within a scale range ($\varepsilon < \varepsilon_{max}$) because of the affinity and periodicity of the fractography textures. Therefore, this characteristic length is associated with the statistical maximum size of the fractography textures. When the measuring scale ε is greater than ε_{max}, the largest and smallest values on the fractography are always included, resulting in $V_{cor}(\varepsilon)=V^2_{max}$.

From the data shown in Table 2, the value of D_{cor} for ductile fracture is the largest, while for brittle intergranular fracture, is the smallest. The results are well in agreement with the irregularity of these fracture surfaces. Fig.1 shows some small dimples are located inside the larger ones. The fractured surface experiences a large amount of ductile deformation and hence dimples at different scales are formed. Such fracture surface has multi-scaling similar structures leading to large ε_{max}, within which variation-correlation follows a lateral growth with the increasing scale. It must be noted that a large ε_{max} does not necessarily mean multi-scaling similar structures, because it is a parameter related to the maximum size of the surface texture. For the brittle intergranular fracture, the fractography is constituted by smooth boundary facets, resulting in a decrease in D_{cor}. The irregularity of the transgranular cleavage fracture is medium in spite of the rather rugged surface because of the smoothness of the surface of each grain.

For the sake of comparison, the curves of height correlation versus scale using height-

height correlation analysis [12-15] are also shown in Fig.1 (e). The height-height correlation function $C(\varepsilon)$ is defined as

$$C(\varepsilon) = < h(\varepsilon) - h(0) >$$

where $h(0)$ is the height of reference point and $h(\varepsilon)$, the height of the points with ε distance from the reference point. The estimated values of the fractal parameters are shown in Table 2. Fractal dimensions from height–height correlation appear the same tendency as those from variation-correlation in characterizing the irregularity of the fracture surfaces. However, the characteristic length ε_{max} in the height-height correlation analysis represents the correlation length [12-15] but not the size of texture. Consequently, the fractal characteristic length ε_{max} in variation-correlation analysis is a more suitable parameter in quantitative fractography because its estimated value is able to represent the maximum size d_{max} of the basic micro-morphology elements. Consider a simple case of brittle intergranular fractography, which simply consists of uniform morphology elements with maximum diameter d_{max}. In variation-correlation analysis, when $d_{max}/2 \leq \varepsilon < d_{max}$, some estimated values of variation in statistical measurement have already reached the value of V_{max}. When $\varepsilon \geq d_{max}$, all the estimated values of variation are V_{max}. Therefore, when $\varepsilon \geq d_{max}/2$, the linearity of the log-log plot becomes worse so that the estimated value of ε_{max} falls into $[d_{max}/2, d_{max}]$ while its fluctuation depends on the given value of M_F. The results in Table 2 hence verify the validity of the analysis. Therefore, the fractal characteristic length can be employed as a statistical measure of the maximum size of texture of fracture surfaces.

4.2 Relationship between fractal dimension and mechanical properties

Fig.2 shows the SEM fractography and the multi-magnification fractal curves for the gray iron. Specimen A has been modified by FeSiCuMo metal stream inoculation. Thus, microstructures of the matrix and graphite structures were improved, leading to an increase in tensile strength σ_b (σ_b=314MPa for sample A and 255MPa for sample B). The elongation for gray iron is almost zero. The size of graphite and graphite-eutectic cells is a

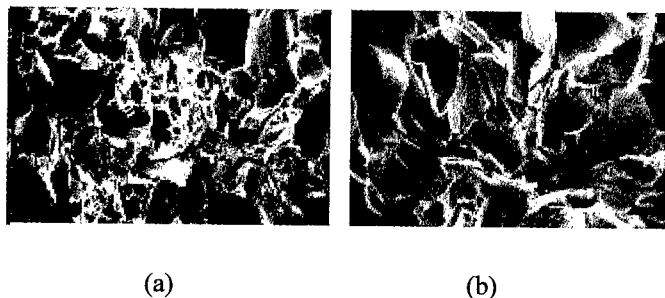
(a)   (b)

Fig.2. SEM fractography of samples A and B of gray iron. (a) SEM fractography of sample A; (b) SEM fractography of sample B. The resolution of the images is 421μm/160pixels.

very important geometric factor. Small grains will lead to multi-scaling fractal structures of the material and the fracture surface becomes more complicated. The cleavage facets of specimen A, which is associated with grain size, are smaller than those of specimen B, leading to the more complicated fractured surface as shown in Fig.2 (a). This implies that the fractal dimensions of specimen A is larger than that of specimen B. Fig.3 is the curve of fractal dimension versus SEM magnification, which verifies this analysis. Fig.3 shows that the fractography cannot be described by a single but rather a series of fractal dimensions. It also reveals that the fractography is not ideally self-affine at all scales of SEM. The transgranular cleavage fracture, for example, shows cleavage facets as dominant morphology at low magnification and river patterns as dominant morphology at high magnification, implying that fractal dimension for a given mode is not universal.
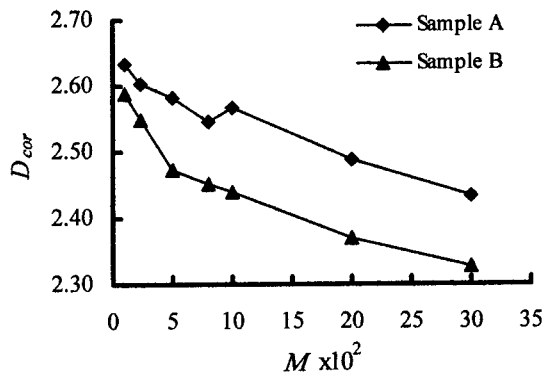
Fig.3. Fractal dimension D_{cor}-magnification M curves of samples A and B.

Fig.4 shows the austenitic dendrite structures of HP26Cr35Ni alloy. Because the specimens were taken from the same region of solidification, the distribution and size of austenitic dendrites of all the specimens are the same even though they have been subjected to different thermal shocks. Their impact toughness, however, is different from each other due to the formation of micro-cracks during the thermal shock treatment. Fractal measurements in Fig.4 (d) show that there is no significant difference in the fractal dimensions of the specimens under different thermal shocks. Therefore, there is no direct correlation between fractal dimension and toughness when the change in toughness is caused by non-geometric factors.

The present experiment has shown that fractal dimensions of fractography at SEM scales are not a single value but rather a series of values for a given fracture mode. This is because of the different dominant patterns of fractography in different range of magnifications. It can be deduced that when the magnification is high enough, the fractal dimension will tend to 2 because the micro-morphology elements approximately become 2D surfaces. The irregular geometric structures of fracture surfaces can reflect the state of the energy absorbed by the fracture surface due to the formation of new surfaces. If the change in toughness is only result of the irregularity of the fracture surface, the fractal can be employed.

(a) (b) (c)

Number of thermal shock

(d)
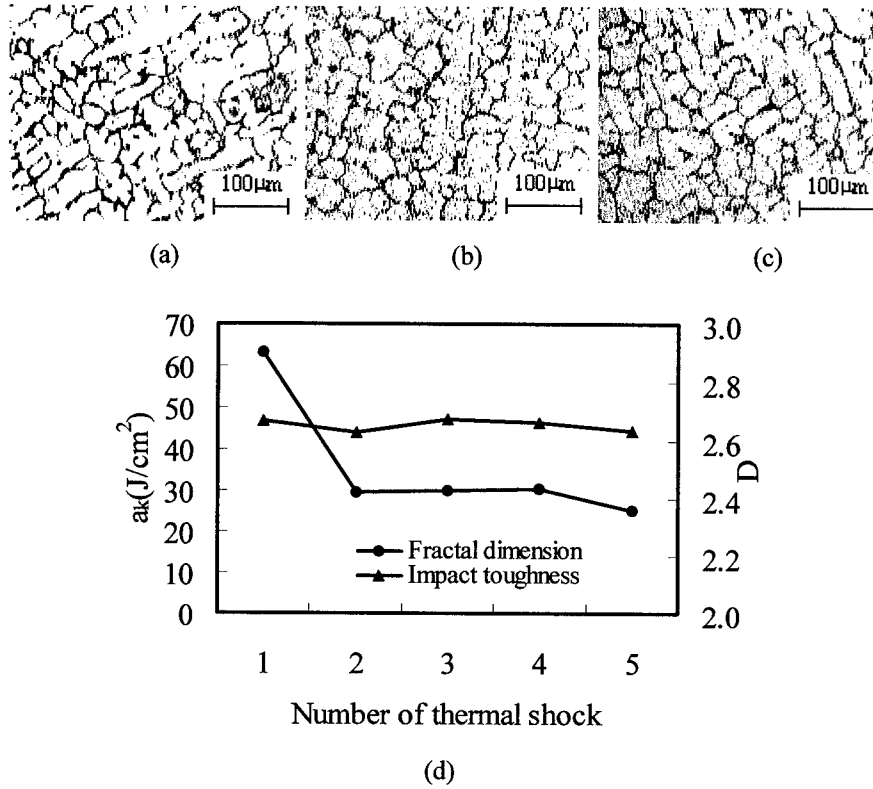
Fig.4. Microstructures, impact toughness and fractal dimension of HP26Cr35Ni alloy specimens subjected to thermal shock of different times. (a) The origin microstructure; (b) microstructure with 10 times of thermal shock; (d) microstructure with 20 times thermal shock; (d) changes in impact toughness and fractal dimension with number of thermal shock.

5 Conclusions

The variation-correlation of fracture surfaces imaged by SEM has been found to follow a power law only within a limited range of scale of $\varepsilon < \varepsilon_{max}$. Beyond this range, the variation-correlation tends to be a constant. The scaling properties can be expressed by

$$V_{cor}(\varepsilon) \propto \begin{cases} \varepsilon^{2\alpha} & \varepsilon << \varepsilon_{max} \\ V_{max}^2 & \varepsilon >> \varepsilon_{max} \end{cases}$$

Fractal dimension of fractography at SEM scales is not a single value but rather a series of values for a given fracture mode. There exists a positive correlation between fractal dimension and toughness of the materials when the change in toughness is caused by geometric factors. However, if the change in toughness is caused by non-geometric factors such as micro-cracks, there is no essential relationship between impact toughness and fractal dimension. For a given fracture mode, fractal dimension is not universal.

References

1. Mandelbrot B. B, Passoja D.E. and Paullay A.J., Fractal Character of Fracture Surfaces of Metals, Nature. 308(1984) pp.721-722.
2. Underwood E. E. and Banerji K., Invited Review Fractal in Fractography, Material Science and Engineering. 80(1986), pp.1-14.
3. Måløy K. J., Hansen A., Lhinrichsen E. and Roux S., Experimental Measurements of the Roughness of Brittle Cracks, Physics Review Letter. 68(1992), pp.213-215.
4. Dubuc B., Quinion J. F., Roques-Carmes C., Tricot C. and Zucker S.W., Evaluating the Fractal Dimension of Profiles, Physics Review. A39(1989), pp.1500-1512.
5. Miller S and Reifenberger R. J., Improved Method for Fractal Analysis Using Scanning Probe Microscopy, Journal o Vacuum Science and Technology. B10(1002), pp. 1203-1207.
6. Denley D. R., Practical Applications of Scanning Tunneling Microscopy, Ultramicroscopy. 33(1990), pp.83-92.
7. Friel J. J. and Pande C. S., A Direct Determination of Fractal Dimension of Fractal Surfaces Using Scanning Electron Microscopy and Stereoscopy, Journal of material Research. 8(1993), pp.100-104.
8. Li J. M., Lu L. Su Y. and Lai M. O., Fractal-Based Description for the Three-Dimensional Surface of Materials, Journal of Applied Physics. 86(1999), pp.2526-2532.
9. Mu Z. Q. and Long C. W., Studies on the Fractal Dimension and Fracture-Toughness of Steel, Journal of physics D: Applied Physics. 21(1988), pp.848-850.
10. Long C.W. and Mu Z. Q., Fractal Dimension Measured with Perimeter-Area Relation and Toughness of Materials, Physical Review. B38(1988), pp.11781-11784.
11. Bouchaud E., Lapasset G. and Planes J., Fractal Dimension of Fractured Surfaces: a Universal Value? Europhysics letters. 13(1990), pp.73-79.
12. Yang H. N., Wang G. C. and Lu T. M., Diffraction from Rough Surface and Dynamic Growth Fronts. Singapore: World Scientific, 1993.
13. Yang H. N., Wang G. C. and Lu T. M., Instability in Low-Temperature Molecular-Beam Epitaxy Growth of Si/Si(111), Physical Review Letters. 73(1994), pp.2348-2351.
14. Palasantzas G. and Krim K., Scanning Tunneling Microscopy Study of the thick Film Limit of Kinetic Roughening, Physical Review Letters. 73(1994), pp.3564-3567.
15. Krim J. and Palasantzas G., Experimental Observations of Self-Affine Scaling and Kinetic Roughening at Sub-micron Lengthscales, International Journal of Modern Physics. B9(1995), pp.599-632.

COARSENING OF FRACTAL INTERFACES

P. STREITENBERGER

Institut für Experimentelle Physik, Abteilung Materialphysik, Otto-von-Guericke-Universität Magdeburg, PSF 4120, D-39016 Magdeburg, Germany

The process of coarsening by curvature driven interface motion of fractal interfaces in two-dimensional space is studied by analytical and numerical methods. A statistical model is presented, which allows an analytical treatment of the main features of coarsening of a fractal interface. For non-conserved motion the interface is described by a statistical distribution function of size scales, which obeys a continuity equation in size space. The solution of the continuity equation yields, for a self-similar initial distribution function, the time development of the interface in terms of a time dependent size distribution function, which exhibits a growing lower characteristic length scale and leads to a power-law decay of the total interface length. The effect of coarsening on the scale of observation is discussed.

1 Introduction

Coarsening is a long-time relaxation phenomenon that is connected with an increase in the characteristic size scale of an unstable two- or multiphase microstructure. The driving force for this scale coarsening is the reduction of the excess free energy associated with the interface between different phases. The interface density $S_V = S/V$, i.e. the interface area S per unit volume V, decreases with time and therefore the characteristic size scale of the interface system given by $l = 1/S_V$ increases, the microstructure coarsens. For sufficiently long times the decrease in the interface density obeys the asymptotic power-law $S_V \approx t^{-\alpha}$, where for sufficiently smooth interfaces the coarsening exponent α only depends on the transport mechanism that drives the interface motion. For curvature driven interface motion (interface limited coarsening), as it is the case for antiphase domain coarsening [1] and normal grain growth [2, 3], the kinetic exponent has the universal value $\alpha = 1/2$ (parabolic growth law). In the case of diffusion limited coarsening the exponent is $\alpha = 1/3$ (Ostwald ripening [2, 4]). For interface limited coarsening the local order-parameter of the phase-ordering dynamics associated with the interface motion is non-conserved, while for diffusion limited coarsening the local order-parameter is conserved [5].

Recently there has been a great deal of interest in the coarsening kinetics of fractal structures. While the generation of fractal patterns far from equilibrium has been widely studied (cf. e.g. [6]), for the reverse process, the decay of a fractal structure, much less work has been done. In complex nonequilibrium systems such as fractal interfaces and clusters the dynamic power law is characterized by a non-trivial coarsening exponent. For the non-conserved curvature driven interface motion it has been shown theoretically as well as experimentally [7 to 10] that initially fractal interfaces exhibit a power law decay with a coarsening exponent α that depends on the fractal dimension. Although similar results has been found, by computer simulation, for the conserved coarsening dynamics of fractal clusters [11 to 13], a complete theoretical description is still lacking for this case [13].

However, also in the case of non-conserved fractal interface coarsening [7 to 10] no exact solution is available because the equation of motion of complex interfaces is highly nonlinear and can, in general, be solved only numerically. Therefore, Toyoki and Honda [7] used a scaling hypothesis to find the asymptotic power law for the coarsening of a non-random self-similar interface. They found for the coarsening exponent the relation $\alpha = (D + 1 - d)/2$, where D is the fractal dimension of the interface and d the Eucleadian dimension of the space. The same result was derived by Orihara and Ishibashi [8], however for a random fractal interface without assuming the self-similar evolution using a diffusion equation for the description of the interface motion. We could confirm the decay law by experimental observation [9, 10] and by a Monte Carlo simulation of the smoothing kinetics of initially fractal two-dimensional grain boundaries [10].

In the present paper a quite different approach is used to describe the coarsening process of a random self-similar interface. The random interface is described by a statistical distribution function of size scales or curvature radii, which obeys a continuity equation in size or curvature space. The solution of the continuity equation yields for non-conserved motion and an initial fractal distribution function the time development of the interface in terms of a time dependent size distribution function, which exhibits a growing lower characteristic length scale. Besides the description of the main features of the coarsening process this approach is also of physical interest because it is an example of how a cutoff of a statistical fractal may emerge in a complex system.

The paper is organized as follows. In order to give an illustrative insight into the coarsening process under consideration, in chapter 2 the dynamic scaling hypothesis and its confirmation (within the range of a lower and upper cutoff) by a Monte Carlo simulation is shortly discussed. In chapter 3 the statistical model is developed which allows for a simple initially random self-similar interface system an analytical solution in terms of a time dependent size distribution function. A summary concludes the paper.

2   Scaling Hypothesis and Monte Carlo Simulation

In general it is not possible to solve the equations of motion of the interface analytically due to the complexity of fractal interfaces in conjunction with the high nonlinearity of the equations of interface motion [7, 8]. Therefore the scaling hypothesis, after which the interface keeps self-similar during temporal evolution, may serve as a simple coarsening scenario [7, 9, 10, 11]. If we restrict ourselves to a plane section of an otherwise three-dimensional interface system the interface density is proportional to the total length L of the boundary line measured on the plane section of area A [9, 10]. For a non-random fractal boundary with a fractal dimension D_S ($1 \le D_S \le 2$), the boundary length obeyes the scaling relation

$$L \approx R_m^{-(D_S-1)}, \tag{1}$$

where R_m is the characteristic length of the structural unit of the boundary corresponding to the lower cutoff length of the fractal. The main effect of coarsening arises from the growth of the lower cutoff due to the time law

$$R_m \approx t^{\frac{1}{\beta+1}}, \qquad (2)$$

where $\beta = 1$ for interface limited and $\beta = 2$ for diffusion limited growth. If we assume that the scaling relation Eq. (1) remains valid also during the relaxation process, the boundary length and therewith the interface density decreases as

$$L \approx t^{-\frac{D_S-1}{\beta+1}}. \qquad (3)$$

For the case of non-conserved interface motion ($\beta = 1$) we have tested the decay law Eq. (3) by performing a Monte Carlo simulation of the thermal relaxation of initially fractal two-dimensional grains [10].

Using the Monte Carlo procedure described by Anderson et. al. [14] for modelling normal grain growth, a single closed grain is mapped onto a discrete square lattice of 512×512 lattice sites, where each lattice site is assigned the orientation number 1 or 2. The interface is defined to lie between two sites of unlike orientations. Taking only an interaction between nearest neighbours into account, the interface motion is simulated by employing a standard Monte Carlo technique in conjunction with an energy criterion for successful transition to the other orientation.
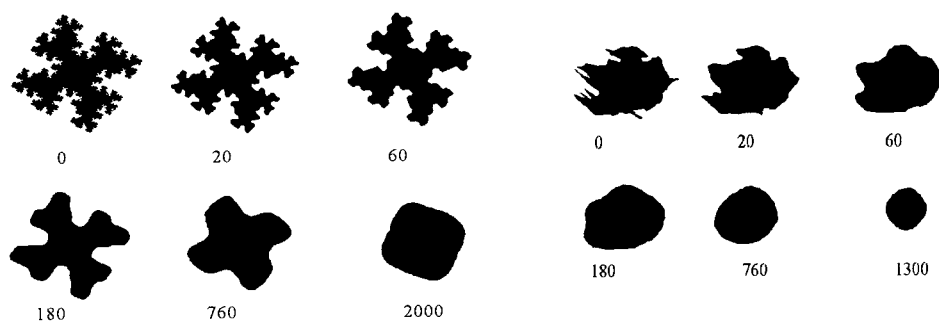
Figure 1: Simulation of the coarsening and shrinking process of an initially non-random fractal ($D_S = 1.5$).The numbers represent the time in Monte Carlo steps.

Figure 2: Same as for Fig. 1 but for an initially random shaped interface ($D_S = 1.22$) from a metallographic section of a deformed Zn specimen.

Figure 1 shows the simulation of the relaxation (and shrinking) process of an initially non-random fractal interface, the so-called square Koch-island ($D_S = 1.5$). The picture shows that the simulation clearly confirms , within a certain range between the lower and upper cutoff, the above scaling assumption that the boundary smoothens out step by step

138

from the higher to about the next lower generation of the initially self-similar fractal pattern. A similar simulation is shown in Fig. 2 starting with an irregularly shaped interface with an initial fractal dimension $D_S = 1.22$, which corresponds to a closed grain boundary from a metallographic section of a deformed Zn specimen (for details see [9, 10]).
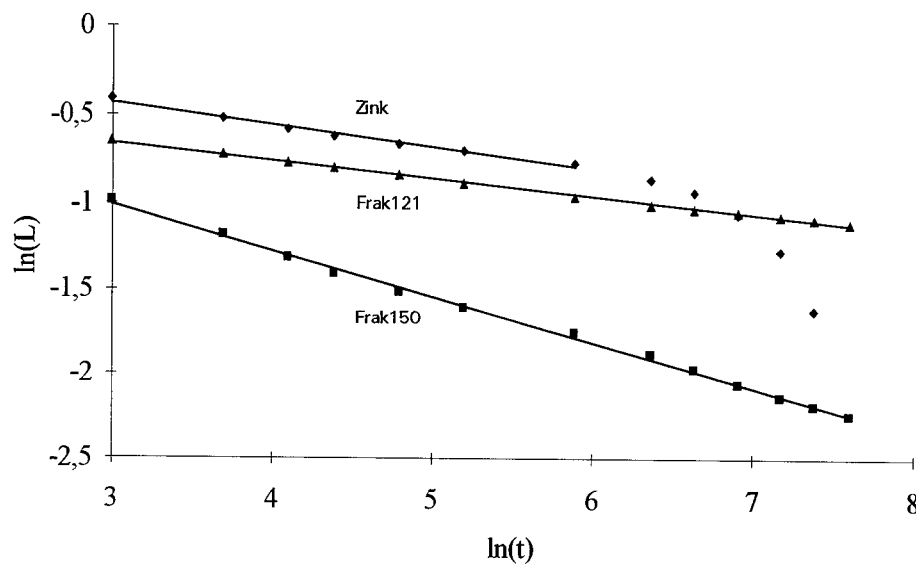
Figure 3: Plot of ln(L) versus ln(t) of the Monte Carlo simulation for three different fractal interfaces. The full lines represent the least-square fit to Eq. (3) ($\beta = 1$) (cf. Table I). L is normalized to the boundary length at $t = 0$.

The log-log plot of the boundary length L (perimeter) of the grains in Figs. 1 and 2 versus simulation time t (measured in Monte Carlo steps) in Fig. 3 shows that the coarsening kinetics, in fact, follows over a wide range the time law Eq. (3) for $\beta = 1$. The fractal dimension derived from the coarsening kinetics, i.e. from the slopes of the full lines in Fig. 3 via the time exponent $(1 - D_S)/2$ in Eq. (3) (for $\beta = 1$), agrees well with the fractal dimension determined by the geometry of the initial interface (Table I). For the random interface (Fig. 2), as time elapses, the crossover from self-similar boundary coarsening to normal grain shrinking kinetics governed by the parabolic time law $L_0^2 - L^2 \approx t$ can be observed in Fig. 3 [10].

In the following paragraph a statistical coarsening model for a much simpler interface configuration is presented, which allows, however, an analytical treatment beyond the above scaling hypothesis of all the main features of self-similar interface coarsening.

	D_S (from the fractal geometry of the initial interface)	D_S (from the coarsening kinetics, Fig. 3 and Eq. (3))
Frak 150: square Koch island	1.50	1.54
Frak 121: modified Koch island	1.21	1.20
Zink: random interface	1.22 (yardstick method)	1.25

3   Statistical Model of Non-Conserved Coarsening of a Self-Similar Interface

In the following considerations we replace the complex fractal boundaries of the preceding paragraph by the somewhat simpler system of a two-dimensional random assembly of non-overlapping and non-interacting closed circular loops of different curvature radii R. For a non-conserved system the equation of motion for a single circular interface with radius R is given by

$$\dot{R} = -kK^{\beta} = -k\frac{1}{R^{\beta}} , \qquad (4)$$

where $\dot{R}$ is the radial velocity, K the curvature and k the mobility of the interface. Again $\beta = 1$ for interface limited growth and $\beta = 2$ for diffusion limited growth. For the description of the time development of the total interface, i.e. of the whole assembly of circular loops, we introduce the size distribution function $F(R,t)$, where $F(R,t)dR$ gives the number of loops per unit area with radius between R and $R + dR$. For a random fractal interface the initial size distribution function is assumed to have the hyperbolic form

$$F(R,0) = F_0(R) = BR^{-D_S-1} , \qquad (5)$$

where D_S is the fractal dimension of the whole assembly and B a constant. Since during coarsening no new loop is created and the only process of annihilation of interfaces is the shrinking of loops according to the growth law Eq. (4) the size distribution function must obey a continuity equation in size space given by

$$\frac{\partial F(R,t)}{\partial t} + \frac{\partial}{\partial R}\left(\dot{R}F(R,t)\right) = 0 . \qquad (6)$$

For a given initial distribution $F(R,0) = F_0(R)$ the solution of the first-order partial differential equation (6) can be written in the form [15]

140

$$F(R,t) = F_0\big(R_0(R,t)\big)\frac{\partial R_0(R,t)}{\partial R}, \tag{7}$$

where $R_0 = R_0(R,t)$ is the solution of the characteristic equation (4) with the integration constant $R_0 = R(t = 0)$. The integration of Eq. (4) yields

$$R_0 = \left(R^{\beta+1} + (\beta+1)kt\right)^{\frac{1}{\beta+1}}. \tag{8}$$

Equations (5) and (8) inserted into Eq. (7) leads to the size distribution function

$$F(R,t) = BR^\beta\left(R^{\beta+1} + (\beta+1)kt\right)^{-\frac{\beta+D_S+1}{\beta+1}}. \tag{9}$$

At $t = 0$ Eq. (9) reduces to the initial distribution Eq. (5). At $t > 0$, in contrast to the
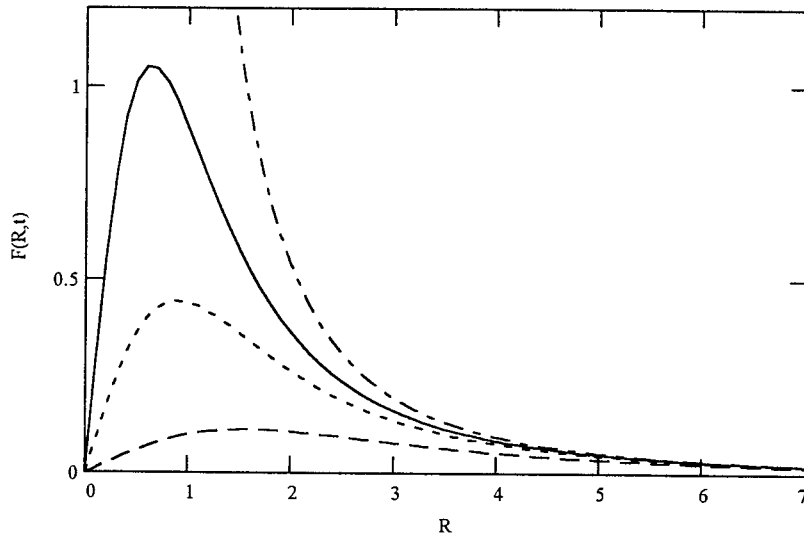
Figure 4: Size distribution function Eq. (9) for $\beta = 1$ and $D_S = 1.5$ at different times t. Dot-dashed line: $t = 0$ (initial distribution Eq. (5)), full line: $t_1 > 0$, dotted line: $t_2 = 2t_1$, dashed line: $t_3 = 6t_1$. R and t are given in terms of arbitrary unit length and unit time, respectively.

initial distribution function, $F(R,t)$ exhibits a maximum and tends towards zero for vanishing radius R (Fig. 4). The location of the maximum defines a natural characteristic

length scale or lower „cutoff" $R_m = R_m(t)$ where scaling breaks down:

$$R_m(t) = \left(\frac{k\beta(\beta+1)t}{D_S + 1} \right)^{\frac{1}{\beta+1}}. \tag{10}$$

In the following discussion we restrict ourselves to interface limited coarsening ($\beta = 1$), where R_m increases with time as

$$R_m(t) = \left(\frac{2kt}{D_S + 1} \right)^{1/2}. \tag{11}$$

The total number of loops decays as

$$N(t) = \int_0^\infty F(R,t)dR = \frac{B}{D_S}(2kt)^{-D_S/2}, \tag{12}$$

which is associated with a decrease in the total length of the interface

$$L(t) = 2\pi \int_0^\infty RF(R,t)dR = 2\pi Bg(0,D_S)(2kt)^{-\frac{D_S-1}{2}}. \tag{13}$$

In this relation $g(0,D_S)$ is defined by the integral

$$g(x,D_S) = \int_x^\infty \xi^2 \left(\xi^2 + 1 \right)^{-\frac{D_S+2}{2}} d\xi. \tag{14}$$

For the average radius we obtain

$$<R> = \frac{L(t)}{N(t)} = 2\pi g(0,D_S)(2kt)^{1/2}, \tag{15}$$

i.e. $<R> \approx R_m(t)$.

The power-law behaviour of the interface length Eq. (13) results in the present approach, without assuming the scaling hypothesis Eq. (1) during the temporal evolution, from the initial fractal distribution function in conjuntion with the growth law Eq.(4) and the continuity equation (6). While for a deterministic fractal the lower cutoff length is simply defined by the lower limit of the geometric building block, for the present statistical

fractal the lower characteristic length scale emerges from the physics of the coarsening process. The emergence of empirical random fractals limited by a lower and upper cutoff has been discussed from a different point of view also in [16].

The present model allows in a simple but instructive way to study the effect of coarsening on the resolution of the interface length by measurement. If we assume that for a certain measuring device the resolution of length scale is restricted to radii $R' \geq R$ the number of loops that can be observed by that device is given by

$$N(R,t) = \int_R^\infty F(R',t)dR' = \frac{B}{D_S}\left(R^2 + 2kt\right)^{-\frac{D_S}{2}} .$$ (16)

The total interface length that can be measured is

$$L(R,t) = 2\pi \int_R^\infty R'F(R',t)dR' = 2\pi B(2kt)^{-\frac{D_S-1}{2}} g(R/(2kt)^{1/2}, D_S) ,$$ (17)

where the function $g(x, D_S)$ is defined by the integral Eq. (14).

Fig. 5 shows in form of a log-log plot the observable interface length Eq. (17) versus
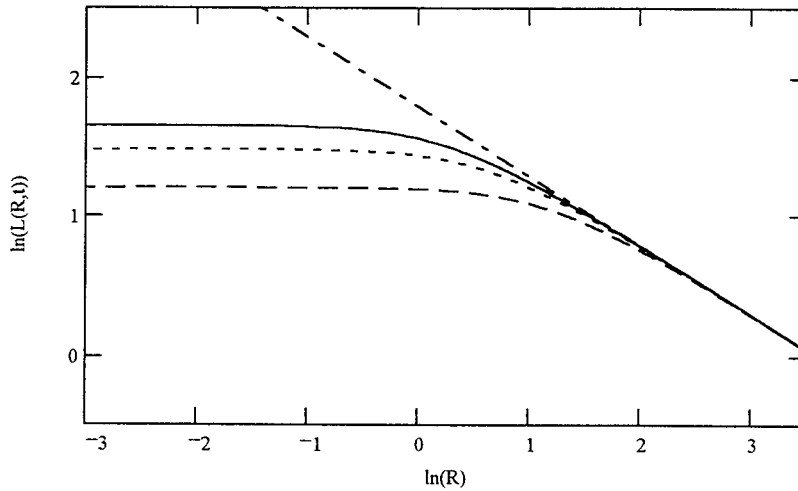
Figure 5: Log-log plot of the interface length Eq. (17) versus R at different times as in Fig. 4.

resolution, that is, in dependence of the segments of scale R that are needed to approximate the interface length (Richardson plot), at different times.

The coarsening, i.e. the growth of the lower characteristic length, leads to a reduction in the available scaling range at smaller R. The width of the crossover region from the eucleadian behaviour at small R to the fractal behaviour at large R becomes broader with increasing time. This behaviour can quantitatively be characterized by the slope of the curves in Fig. 5, or alternatively, by the effective fractal dimension [17]

$$D_{eff}(R,t) = 1 - \frac{d \ln L(R,t)}{d \ln R},$$

(18)

which depends on the scale of observation R. $D_{eff} = D_{eff}(R,t)$ shows a crossover from $D_{eff} = 1$ at small R (high resolution) to $D_{eff} = D_S$ at large R (low resolution) (Fig. 6). For later times corresponding to a more relaxed interface, D_{eff} approaches D_S at larger R. This may have an effect on the empirical fractal dimension of the interface, which is obtained by application of an empirical observation method, as the yardstick or box counting method, at a fixed coarsening state.
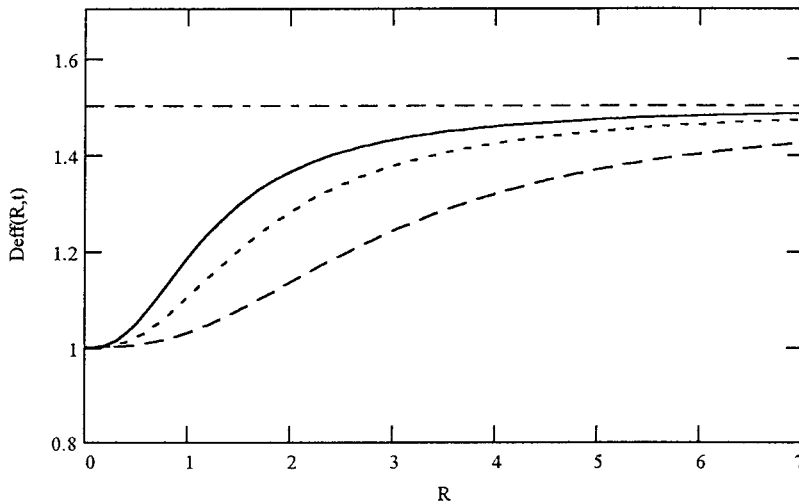
Figure 6: Scale dependent effective fractal dimension Eq. (18) at different times as in Fig. 4.

If we consider D_{eff} as an estimate for the empirical fractal dimension and if we take into account that for a sufficiently coarsened interface D_{eff} may remain well below the asymptotic value D_S in Fig. 6 because the usable scale of R is limited by the existence of an upper cutoff defined by the sample size (e.g. by the grain size in Fig. 1), it follows that the empirical fractal dimension decreases during coarsening. Hence, for each coarsening state of the grain in Fig. 1, for example, a somewhat lower empirical fractal dimension

than for the previous state might be found. In contrast, the corresponding coarsening kinetics, as represented by the slope of the straight lines in Fig. 3, is uniquely determined by the „initial" fractal dimension D_S that is inherent to the initial self-similar interface or, as in the present statistical model, to the initial distribution function Eq. (5).

Finally, it is worth noting that the functional form of the fractal equation (17) is quite different from Rigaut's empirical interpolation Ansatz [18] for concave log-log plots. Rigaut's interpolation formula has the form

$$L = c_1 (1 + c_2 R^{D_S - 1})^{-1}, \tag{19}$$

which tends in a log-log plot as Fig. 5 towards the horizontal maximum at high resolution and towards the ideal fractal at low resolution. From the limiting behaviour of Eq. (17) for small and large R one finds $c_1 = L(t)$ and $c_2 = (D_S - 1)L(t) / (2\pi B)$, respectively, where $L(t)$ is given by Eq. (13). A quantitative comparison shows that this interpolation formula represents only a poor approximation to the exact fractal equation (17) in the crossover region.

4 Summary

After a discussion of the dynamic scaling behaviour of self-similar interfaces as it follows from analytical scaling assumptions and numerical Monte Carlo simulations a statistical model was presented, which shows all the main features of coarsening of fractal interfaces. The coarsening kinetics of a two-dimensional non-conserved interface system consisting of a random assembly of non-interacting circular loops could be described by a size distribution function obeying a continuity equation in size space. For a self-similar initial distribution the solution of the continuity equation yielded a time dependent size distribution function, which exhibts a lower characteristic length scale where scaling breaks down. Physically this emerges from the curvature driven relaxational motion of the interfaces leading to parabolic growth of the characteristic length scale. The total interface length shows the characteristic power-law decay, where the coarsening exponent depends on the inherent fractal dimension of the initial interface. The model allowed us to calculate the interface length as a function of resolution, which led us to a discussion of the effect of coarsening on the scale of observation.

References

1. S. M Allen and J. W. Cahn, *Acta metall.* **27**, 1085 (1979).
2. W. W. Mullins and J. Vinals, *Acta metall.* **37**, 991 (1989).
3. H. Weiland, B. L. Adams, and A. D. Rollett, *Grain Growth in Polycrystalline Material III* (The Minerals, Metals & Materials Society, Warrendale, 1998).
4. I. M. Lifshitz and V. V. Slyozov, *J. Phys. Chem. Solids* **19**, 35 (1961).
5. A. J. Bray, *Adv. Phys.* **43**, 357 (1994).
6. P. Meakin, *Fractals, Scaling and Growth Far From Equilibrium* (Cambridge University Press, Cambridge, 1998).

7. H. Toyoki and K. Honda, *Phys. Lett.* **111**, 367 (1985).

8. H. Orihara and Y. Ishibashi, *J. Phys. Soc. Japan* **56**, 2340 (1987).

9. P. Streitenberger, D. Förster, G. Kolbe, and P. Veit, *Scripta metall. mater.* **33**, 541 (1995); *Scripta mater.* **34**, 111 (1996).

10. P. Streitenberger, D. Förster, and P. Veit, *Fractals* **5**, Suppl., 5 (1997).

11. T. Irisawa, M. Uwaha, and Y. Saito, *Europhys. Lett.* **30**, 139 (1995).

12. R. Thouy, N. Olivi-Tran, and R. Jullien, *Phys. Rev.* **B 56**, 5321 (1997).

13. M. Conti, B. Meerson, and P. V. Sasorov, *Phys. Rev. Lett.* **80**, 4693 (1998).

14. M. P. Anderson, D. J. Srolovitz, G. S. Grest, and P. S. Sahni, *Acta metall.* **32**, 783 (1984)

15. V. V. Slyozov, J. Schmelzer, and J. Möller, *J. Cryst. Growth* **132**, 419 (1993).

16. O. Malcai, D. A. Lidar, O. Biham, and D. Avnir, *Phys. Rev.* **E 56**, 2817 (1997).

17. H. Takayasu, *Fractals in the Physical Sciences* (Manchester University Press, Manchester, 1990), p. 141.

18. J. P. Rigaut, *J. Microsc.* **133**, 41 (1983).

A FRACTAL MODEL OF OCEAN SURFACE SUPERDIFFUSION

P.S. ADDISON, B. QU

*Civil Engineering Group, School of the Built Environment, Napier
University,10 Colinton Road, Edinburgh, EH10 5DT, Scotland, UK.,*

G. PENDER

Dept. of Civil Engineering, Glasgow University, Glasgow, Scotland, UK.

The transport of surface pollutants in the coastal zone is modelled using a modified particle tracking diffusion model. The new model uses fractional Brownian motion (fBm) functions to produce superdiffusive spreading of the synthesised pollutant clouds. The model is tested on a numerical model of a coastal bay recirculation zone.

1 Introduction

The spread of pollutants in the environment is a topic of much research. Many environmental pollutants are transported via turbulent fluids (in rivers, the oceans and the atmosphere). The overall spreading of the pollutants in these turbulent flowfields is generally non-Fickian (in fact superdiffusive) over some, if not all, length scales under consideration. Traditional methods of modelling diffusive processes in the environment rely on either solving an advection-diffusion equation on a computational grid or using a particle-tracking technique[1-3]. Both methods lead to solutions which are Fickian in nature. The spatial correlations which exist over significantly large scales result in a Lagrangian memory effects within the flow field. Recently the authors have developed a particle-tracking technique that produces non-Fickian diffusion[4-6]. The technique works by incorporating fractional Brownian motion (fBm) trajectories within the particle tracking model. The work is motivated by the recent discovery that ocean surface drifter trajectories are fractal in nature[7-9]. In this paper previous work is expanded and the simulation of coastal diffusion in a numerically generated bay flow model is conducted.

2 The Non-Fickian Particle Tracking Model

Particle tracking models work by releasing a large number of massless marked particles into a known flow-field. These are then diffused due to a Fickian diffusion model and advected according to the spatial distribution of the flow-field. Each particle represents a portion of the mass of the contaminant, and the ensemble particle cloud can be converted into a spatial concentration distribution. The authors generate non-Fickian diffusive behaviour of particle clouds within a particle tracking diffusion model by specifying each particle path as a fractional Brownian motion (fBm). A generalisation of Brownian motion, fBm is a random fractal function defined by Mandelbrot and Van Ness[10] which exhibits long term correlation over all scales. In this section we briefly give the algorithm

148

for the generation of fBm, more specific details of the model are given by the authors elsewhere, see for example[11-13]. Fractional Brownian motion is defined as

$$B_H(t) = \frac{1}{\Gamma(H + \frac{1}{2})}\left[\int_{-\infty}^{0}\left[(t-t')^{H-\frac{1}{2}} - (-t')^{H-\frac{1}{2}}\right].R(t').dt' + \int_{0}^{t}.(t-t')^{H-\frac{1}{2}}.R(t').dt'\right], \quad (1)$$

where Γ is the gamma function and $R(t)$ is a continuous white noise function. A discrete approximation to Eq. (1) may be generated in two steps: Firstly, incremental steps in the fBm walk are calculated using

$$B_H(t_i) - B_H(t_{i-1}) = \frac{1}{\Gamma(H+\frac{1}{2})}\left[\left[\sum_{j=i-M}^{i-2}\left((i-j)^{H-\frac{1}{2}} - (i-j-1)^{H-\frac{1}{2}}\right).R(t_j)\right] + R(t_{i-1})\right], \quad (2a)$$

where M is a finite memory; and $R(t_j)$ are discrete random numbers with a Gaussian distribution of known standard deviation and zero mean. (Note that simpler distributions may be used.) These incremental steps are known as fractional Gaussian noise (fGn) - generalisations of white noise. The fGn is then summed to generate fBm at discrete times $t_i (=\Delta t \times i)$

$$B_H(t_i) = \sum_{k=1}^{i}\left[B_H(t_k) - B_H(t_{k-1})\right]. \quad (2b)$$

Eqs. (2a) and (2b) may then be used to specify the displacement-time behaviour of individual particles in each spatial dimension of the diffusive problem under investigation. This requires an independent realisation of the fBm for each spatial co-ordinate. Carrying this out for each particle in a diffusing cloud results in scaling of standard deviation of the cloud, σ_c, of the form

$$\sigma_c = \left(2.K_f.t\right)^{H} \quad (3)$$

(here we do not include advection), where K_f is the fractal (anomalous) diffusion coefficient and H is the Hurst exponent[14].
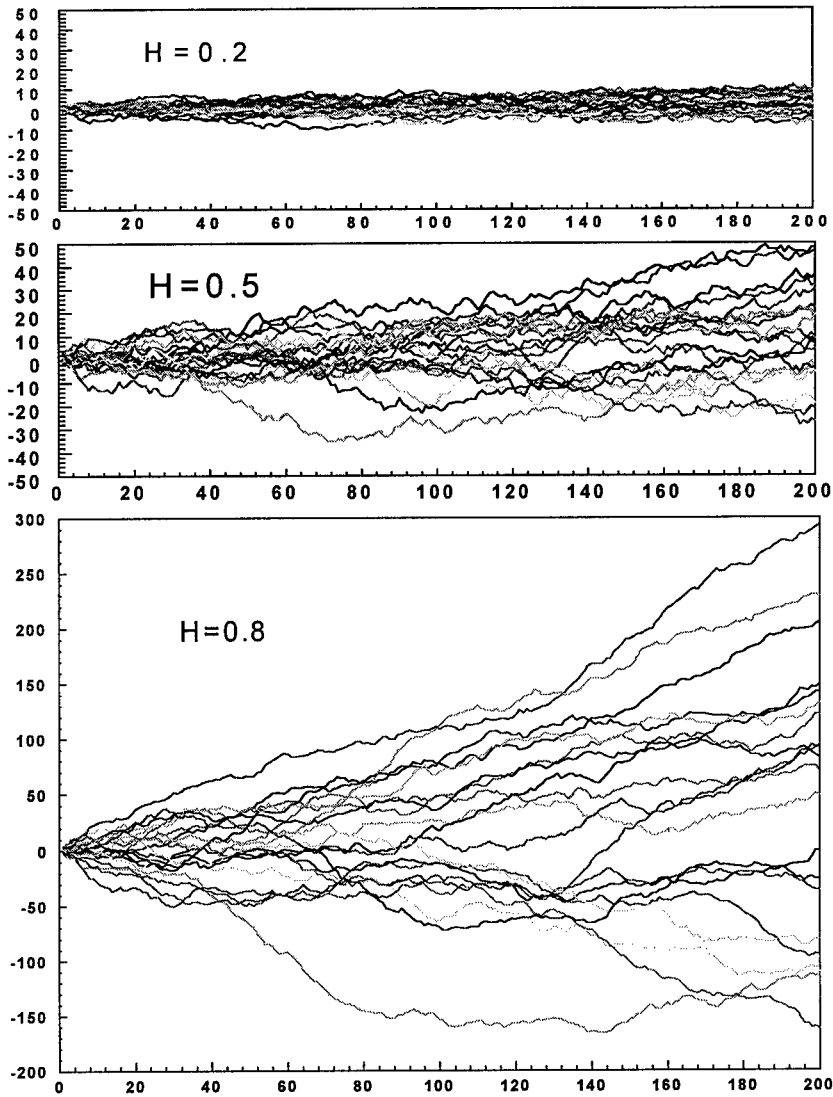
Figure 1: Clouds of 20 fBm Traces H = 0.2, 0.5, 0.8. (Note: Different Scales on Vertical Axes.).

Fig. 1 shows the diffusion of clouds of fBm (H=0.2, 0.5 and 0.8) from a point source. The respective subdiffusive and superdiffusive nature of the H=0.2 and 0.8 fBm is evident in the plot. In order to generate the fBm in two (or more) spatial directions an fBm trace is used for each independent co-ordinate. Figure 2 contains 2D fBm spatial trajectories for single particles. Again the subdiffusive nature of the H=0.2 path and superdiffusive nature of the H=0.8 path can be seen. The particle tracking model works by releasing many 1000's of particles utilising their ensemble spreading behaviour.
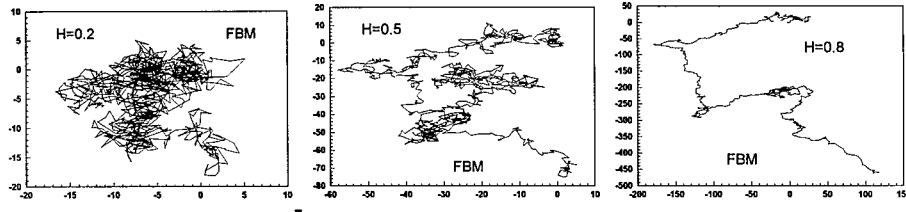
Figure 2: Two-Dimensional fBm trajectories.

The finite memory, *M*, of the synthesised fBm is an important factor that must be taken account of when generating the fBm using the above method. Memories that are too low result in a poor representation of the fBm over the problem timescale. Figure 3 shows the effect of low memory on the resulting power law of the diffusing cloud. In the figure a 100 step fBm is generated using memories from 100 to 2000 steps, i.e. from one times the length of each component trace to twenty times the length of each trace. The gradual straightening of the curve is evident from the plot indicating that the synthesised fBm realises the scaling power law (Eq. 3) as the memory increases. The authors have found that a memory of approximately 5 times the problem lengthscale is suitable for most practical applications.
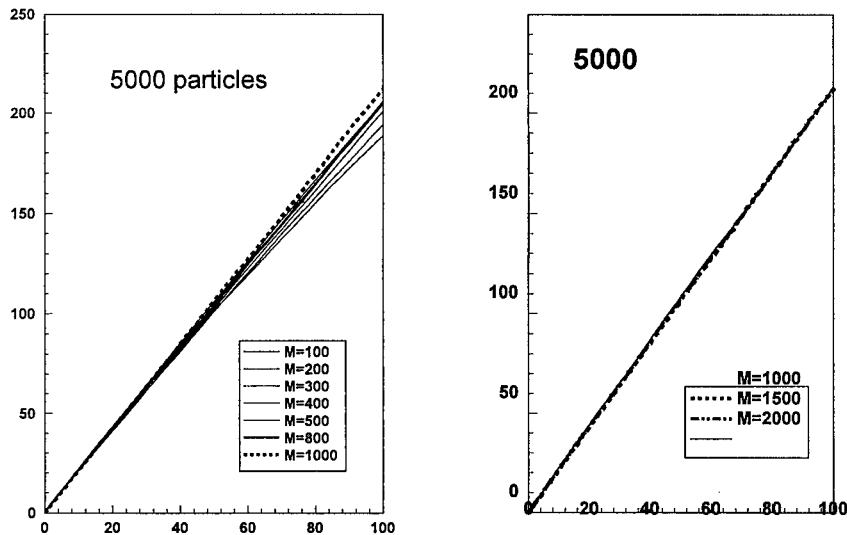
Figure 3: $\sigma^{\frac{1}{H}}$ versus time for various values of memory *M*.

3 Dispersion in a Coastal Bay

Fig. 4 shows a numerically generated surface velocity field for a coastal bay with a North-South flow. This main flow causes a recirculation zone within the bay itself. In such complex flow-fields the relative change in the velocity vectors (magnitude and direction)

can cause additional dispersion of the pollutant cloud to take place. This effect is known as shear dispersion.
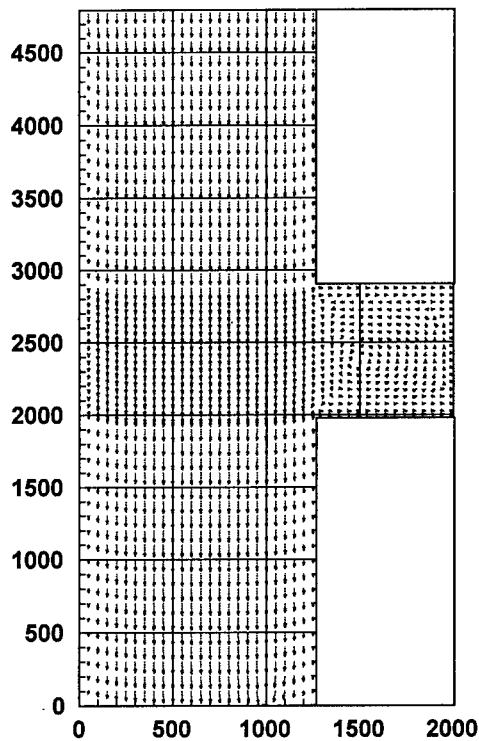
Figure 4: Coastal Bay Model, Surface Velocity Vector Plot.

Fig. 5 shows the comparison of two synthesised pollutant clouds, released in the bay, as they are dispersed over time. The left hand plots for each pair contains the isoconcentration contours of a Fickian cloud (H=0.5) and the right hand plots for a superdiffusive (H=0.8) cloud. Within the bay itself we see that the non-Fickian cloud spreads much more rapidly than the (traditional) Fickian cloud. This rapid spreading can result in areas being affected that would otherwise escape contamination with the Fickian model. In addition, lower concentrations are reached much sooner with the superdiffusive cloud but a larger area is affected. This has implications for the modelled biological impact of the pollutant on the environment, especially for those organisms whose mortality depends on a threshold concentration of contaminant.

152

Figure 5: Comparison of Fickian ($H = 0.5$) and Non-Fickian ($H = 0.8$) Contaminant Clouds Released from (1300, 2800) at Time Zero. Plots show location of cloud after 1, 2, 3, 4, 7, 10 hours.

4 Concluding Remarks

The non-Fickian particle tracking diffusion model developed by the authors allows for a more flexible approach than currently available to the modelling of contaminant transport within turbulent flowfields. The model has been illustrated using a coastal bay flow-field, however, it has general applications to non-Fickian diffusive processes in one, two or three dimensions. (The authors have developed a version for subsurface diffusion through material of variable hydraulic conductivities[12,15]). In the study, detailed above, the effects of shear dispersion, wind shear and tidal motions were neglected. In a fully working model these important dispersive processes would have to be taken account of.

One drawback of the model is that H is restricted to vary between 0 and 1. However, it is known that diffusive processes on the open ocean (i.e. far from boundaries) can scale with Hurst exponents up to 1.5. It is possible that by incorporating shear dispersion, super-diffusive processes may be realised with effective exponents above unity. Another drawback of the model is that the fBm generation method, given by Eqs. (2a) and (2b), requires a heavy computational effort. This is because each fGn requires a summation over M time steps, where the memory M must be of the order of (at least) five times the duration of the problem under investigation to ensure accurate non-Fickian statistics over the problem timescale. The authors have recently developed faster spectral methods of generating fBm based on the work of Yin[16]. Current work by the authors concentrates on the both the effect of shear dispersion on fBm dispersive processes and the implementation of the fBm particle tracking model to a fully three dimensional numerical model with experimental verification.

References

1. C. M. Allen, 'Numerical simulation of contaminant dispersion in estuary flows', *Proc. Roy. Soc. Lond. A*, 381, 179-192 (1992).
2. C. M. Allen, 'Particle tracking models for pollutant dispersion', in D. G. Farmer and N. J. Rycroft (eds), *Computer Modelling in the Environmental Sciences*, IMA/Clarendon, Oxford, 1991.
3. J. R. Hunter, P. D. Craig and H. E. Philips, 'On the use of random walk models with spatially variable diffusivity', *J. Comput. Phys.* 106, 366-376 (1993).
4. P. S. Addison, 'A method for modelling dispersion dynamics in coastal waters using fractional Brownian motion', *IAHR J. Hydraulic Research* 34, 549-561 (1996).
5. P. S. Addison and B. Qu, 'Non-Fickian random walk diffusion models', in'Environmental and Coastal Hydraulics: Protecting the Aquatic Habitat', *Proceedings of Theme B, Volume 1, XXVII IAHR Congress, 10-15 August, San Francisco, 1997*. ASCE, 45-50. (1997)
6. P. S. Addison, B. Qu, A. Nisbet, G. Pender, 'A non-Fickian, particle-tracking diffusion model based on fractional Brownian motion', *Int. J. Num. Meth. Fluids* 25, 1373-1384 (1997)
7. A. R. Osborne, A. D. Kirwan, A. Provenzale, and L. Bergamasco, 'Fractal Drifter Trajectories in the Kuroshio Extension', *Tellus* 41A, 416-435. (1989)
8. B. G. Sanderson, A. Goulding, and A. Okubo, 'The Fractal Dimension of Relative Lagrangian Motion', *Tellus* 42A, 550-556. (1990)
9. B. G. Sanderson and D. A. Booth, 'The Fractal Dimension of Drifter Trajectories and Estimates of Horizontal Eddy-Diffusivity', *Tellus* 43A, 334-349. (1991)

154

10. B. B. Mandelbrot and J. W. Van Ness, 'Fractional Brownian Motions, Fractional Noises and Applications', *SIAM Review* 10, 422-437 (1968).
11. P. S. Addison, 'Towards a non-Fickian Diffusion Model using Fractal Geometric Techniques', *ASCE J. of Hydraul. Engng*, 124(9), 877-879. (1998).
12. P. S. Addison, B. Qu, A. S. Ndumu and I. C. Pyrah, 'A particle-tracking model for non-Fickian subsurface diffusion', *Math. Geology* 30 (6), 695-716 (1998)
13. P. S. Addison, *Fractals and Chaos: An Illustrated Course*, Institute of Physics Publishing, Bristol, 1997.
14. H. E. Hurst, 'Long Term Storage Capacity of Reservoirs', *Trans. ASCE* 116, 770-808, (1951).
15. P. S. Addison and A. S. Ndumu, 'Engineering Applications of Fractional Brownian Motion: Self-Similar and Self-Affine Random Processes', *Fractals:* 7(2), 151-157. (1999).
16. Z. -M. Yin, 'New Methods for Simulation of Fractional Brownian Motion', *Jour. of Comp. Physics*, 127, 66-72 (1996).

STOCHASTIC SUBSURFACE FLOW AND TRANSPORT IN FRACTAL CONDUCTIVITY FIELDS

ALBERTO S. NDUMU, PAUL S. ADDISON

Civil Engineering Group, School of the Built Environment, Napier University,

10 Colinton Road, Edinburgh, EH10 5DT, Scotland, UK.

Monte Carlo simulations of subsurface flow and contaminant transport of a non-reactive solute plume by steady-state flow with a uniform velocity were performed in a two-dimensional synthetic heterogeneous porous media whose hydraulic conductivity is non-stationary and described by multi-scale fractional Brownian motion. Analysis of the flow and transport results indicates that the longitudinal velocity variance is nearly constant in the longitudinal direction while in the transverse direction it assumes a parabolic shape. The velocity variance is maximum at the impervious boundaries and decreases in transverse direction with distance from the boundaries reaching the minimum value at the domain centre. We observe that the particle displacement covariance is anomalous or non-Fickian at all times t in the dispersion process irrespective of the Hurst exponent H and grows temporally faster than linearly.

1 Introduction

Natural variability of subsurface (i.e., hydrologic and geologic systems) hydraulic properties is the main factor controlling the flow and spreading of contaminants in subsurface porous media. The primary physical property which exhibits large-scale natural spatial variability is the hydraulic conductivity. One important source of error in numerical models to simulate flow and transport in the subsurface stems from lack of knowledge concerning the spatial variability of hydraulic conductivity. In practice, we typically have sparse data which is inadequate to completely describe the spatial variability. This is a consequence of the high data collection cost required to fully characterise a given subsurface porous medium. Hence, the variation of subsurface hydraulic conductivity cannot generally be described in all its detail and is therefore uncertain[1].

One response to the problem of the uncertainty of subsurface hydraulic conductivity has been an interest in stochastic methods, which provide a formal framework for the treatment of uncertainty. Spatial variability of subsurface properties is such that their unique, deterministic, description is not feasible[2], and this is formally recognised in the stochastic approach, in which subsurface hydraulic conductivity is commonly regarded as a random scalar field (RSF).

Stochastic theories involve the description of the local porous medium structure using a statistical model that requires a small number of parameters to be identified from field measurements. The detailed spatial distribution of logconductivity is conventionally reduced to a few statistical parameters, for example, covariance function, mean, variance, and correlation length. It thus follows that the outputs from stochastic models are probabilistic, characterised, for example, by the statistical moments or the full probability density function of the variable of interest.

In this study, we perform Monte Carlo experiments by generating a large number of independent random conductivity realisations with a fractal semivariogram function, solving the stochastic boundary value problem by repetitively solving a set of

deterministic flow problems, each of which is an equally probable representation of the response of the real heterogeneous medium. The flow problem is solved by employing a block-centred finite difference scheme (e.g., Smith and Freeze[3]) to obtain subsurface velocity fields. A random walk particle tracking algorithm is used to solve the transport equation (e.g.,[4-7]) using realised velocity fields. Realised solutions are averaged to get fluctuating log conductivity semivariograms, velocity variances, particle displacement covariance and mean solute concentrations.

2 Random Field Model

In the Monte Carlo simulation of subsurface flow and transport, the first step is the choice of the statistical model that represents the medium heterogeneity, mainly the log hydraulic conductivity, $\ln K(\mathbf{x})$. Most stochastic theoretical and numerical modelling approaches (see e.g.[7-9]) assume that the log hydraulic conductivity $Y(\mathbf{x}) = \ln K(\mathbf{x})$ where $K(\mathbf{x})$ is the local hydraulic conductivity of the subsurface porous medium and $\mathbf{x}$ is the vector of spatial co-ordinates, as stationary random scalar field, normally distributed and characterised by a constant mean $\langle Y \rangle$ and isotropic exponential covariance functions e.g.,

$C_Y(r) = \sigma_Y^2 e^{-r/\lambda}$ (with $r = |\mathbf{r}|$, where $\mathbf{r}$ is the two-point isotropic lag, σ_Y^2 is the variance of the distribution, and λ the finite integral scale). The justification for using this model is based on geostatistical data obtained from several sites (e.g.,[10]). In addition, the macrodispersivity measured from tracer tests agree approximately well with the stochastic theory results based on the exponential model. Zhan and Wheatcraft[11] argue that, field hydraulic conductivity measurements are limited, have large uncertainty and have been carried out for relatively small spatial scales (at most a kilometre). If the measurements are available at much larger spatial scales, conductivity values may remain correlated at larger scales and may give rise to *non-stationary fractal* or *self-similar* distributions[12-14] with infinite correlation length. Neuman[15] proposes a spatial scaling assumption with a semivariogram function given in (1) to describe such a distribution as

$$\gamma_Y(r) = \frac{1}{2}\left\langle [Y(\mathbf{x} + \mathbf{r}) - Y(\mathbf{x})]^2 \right\rangle = \gamma_0 (r/\lambda)^{2H}, \tag{1}$$

where λ is a reference scale, γ_0 is a constant variance parameter, and the scaling exponent H is the Hurst exponent[16]. Equation (1) has been demonstrated to be valid for self-affine stochastic processes over the broad range $0 < H < 1$[24]. The exponent H is associated with the fractal dimension $D_S = E + 1 - H$, where E is the space dimensionality. For $0 < H < 1$ the variance and integral scale of the field Y grow infinitely, thus describing a medium with spatially evolving heterogeneity. The semivariogram, given by (1), shows a continuous growth with distance. The presence of more than one reference scale of variability has an influence on fluid flow and contaminant migration[6]. Spatially evolving scale formations are characterised by a variance which keeps growing with the domain size. For a bounded domain, the order of magnitude of the variance is given from (1) by the following:

$$\sigma_Y^2 = \gamma_0 R^{2H} , \qquad (2)$$

where R is the characteristic dimension of the domain (upper fractal cut-off scale). Hence, the log conductivity integral scale λ shows the following scaling property:

$$\lambda \propto \alpha R \qquad (3)$$

where α is a constant that depends on H. For independent realisations of the log conductivity field, the variance of the random processes varies because of lack of stationarity. However, averaging over many realisations, the order of magnitude of the variance is controlled by the above relationship. The spatial variability of the field is controlled by H while, for a given H, the contrast existing between the actual values of log conductivity Y is controlled by γ_0. Increasing γ_0, keeping H constant, amplifies the contrast between block conductivites. To perform hydraulic conductivity simulations for different H values, the same order of magnitude of the fluctuations needs to be maintained. This was needed in order to obtain realistic fluctuations for the case $H > 1/2$. From a practical point of view we impose in all cases the following condition:

$$\gamma_0 R^{2H} = \sigma_Y^2 , \qquad (4)$$

where σ_Y^2 is the variance of the field at the maximum size of the modelled domain (upper cut-off scale). This means that once we fix the value γ_0 for the case $H < 1/2$, say $[\gamma_0]_1$ for $H = H_1$, the coefficient $[\gamma_0]_2$ for the case $H = H_2 > 1/2$ becomes:

$$[\gamma_0]_2 = \frac{[\gamma_0]_1 R^{2H_1}}{R^{2H_2}} . \qquad (5)$$

Field measurements of hydraulic conductivity have indicated an approximately log-normal distribution[10]. The same distribution has therefore been adopted for the simulations generated in this study. Essentially, a normally distributed log conductivity field $Y(x,y)$ of stationary increments and isotropic semivariogram given in (1) is generated over a two-dimensional domain, the hydraulic conductivity field is obtained by the transformation

$$K(x,y) = \exp[Y(x,y)] , \qquad (6)$$

where $K(x,y)$ is the conductivity assigned to the point x,y of the domain. The turning band method[19,20] was used to simulate the hydraulic conductivity fields. Using this approach we do not require to filter the wavelengths larger than the field dimension (fractal cut-off scale), from the Y spectrum. The larger scale of variability is limited by the field dimension, and the cut-off is introduced by the fact that the generation is performed in a finite domain similar to the work of Bellin et al.,[6]. Since the log conductivity field Y is non-stationary, the mean value of Y is maintained constant in each realisation of the field Y, by conditioning the field to a given constant value[21]. This

158

translates to conditioning the mean velocity field to a given constant value. Figure (1a) and (1b) shows two realised fractal hydraulic conductivity fields generated using the turning bands method. The simulated fields pertain to $H = 0.3$, and $H = 0.8$ respectively. These fields show dramatically different behaviour. When $H > 1/2$, Y is positively correlated and the fields shows smooth variations and large-scale persistence[22] of the positive and negative values, Fig. 1b. Conductivity values tend to cluster above or below the mean for quite some distance before they change to the other side of the mean. When $H < 1/2$, Y is negatively correlated and leads to less persistence within the conductivity values, and hence Y varies erratically, Fig.1a.

(a) $H = 0.3$ (b) $H = 0.8$
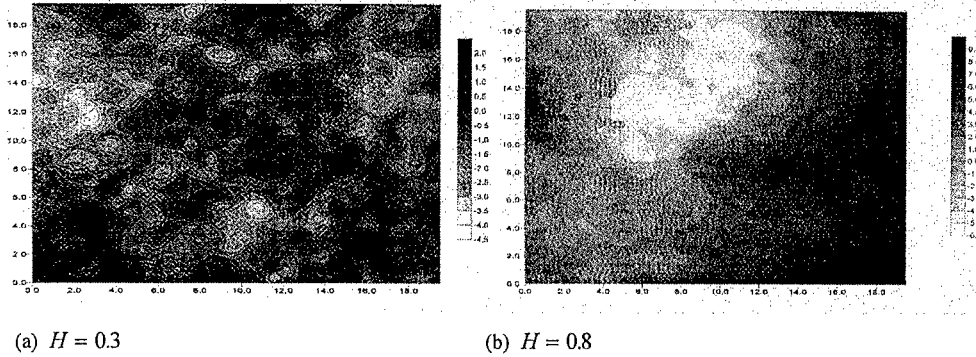
Figure 1: Fractal Hydraulic Conductivity Fields Simulated by Turning Bands Method.

To address model accuracy, we consider the following statistic:

$$\gamma_Y^s(\mathbf{r}) = \frac{1}{2N} \sum_{j=1}^{N} \left[Y(\mathbf{x}_j) - Y(\mathbf{x}_j - \mathbf{r}) \right]^2 \tag{7}$$

where N denotes the number of pairs of log hydraulic conductivity Y with separation equal to r. $\gamma_Y^s(\mathbf{r})$ is the semivariogram obtained by spatial averaging all the equi-distant pairs in a single realisation over a single replicate and assuming stationarity. The reconstructed expected value of the log conductivity semivariograms for $H = 0.3, 0.8$, are shown in Figure (2a) and (2b) respectively. The expected values are computed in a Monte Carlo sense averaging over a number of MC independent realisations as

$$\gamma_Y^{MC}(\mathbf{r}) = \left\langle \gamma_Y^s(\mathbf{r}) \right\rangle = \frac{1}{MC} \sum_{j=1}^{MC} \gamma_Y^{s,(j)}(\mathbf{r}) \tag{8}$$

according to the following convergence criteria[6]

$$\left\{ \frac{1}{M} \sum_{k=1}^{M} \left[\gamma_Y^{MC}(\mathbf{r}_k) - \gamma_Y^{MC-1}(\mathbf{r}_k) \right]^2 \right\}^{1/2} \leq \varepsilon, \tag{9}$$

where r_k is the distance of the two point lag along the North-South (NS) and East-West (EW) directions of the modelled domain and M is the number of points used to discretise the semivariogram. The simulated and theoretical semivariograms (Eq. (1)) along NS (squares) and EW (triangles) directions for different H values agree well in both cases in Fig. 2 although the accuracy deteriorates with lag. The plotted semivariograms shows that the variance grows infinitely with lag distance for a medium described by a semivariogram given in Eq. 1.
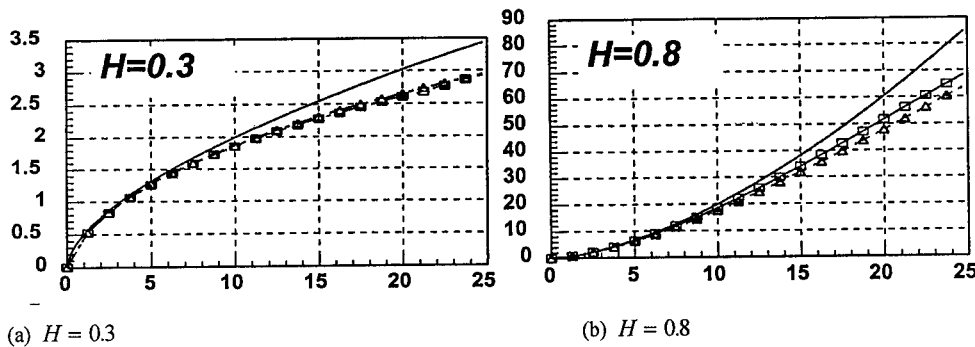

(a) $H = 0.3$ (b) $H = 0.8$

Figure 2: Simulated $\gamma_Y^{MC}(r)$ and theoretical $\gamma_Y(r) = a(r/\lambda)^{2H}$ semivariograms along NS and EW for fields in Figure (1).

3 Stochastic Finite Difference Flow Model

Steady-state confined flow of groundwater in a two-dimensional, horizontal, saturated, imcompressible porous medium with physical heterogeneity represented by simulated spatially varying fractal conductivity fields is considered in this study. Flow simulations were performed in a Monte Carlo manner for fractal fields of various H values. The simulation domain is a square domain where constant hydraulic heads are assigned at the left and right boundary and the no-flow condition on the top and bottom to create a uniform mean velocity $\bar{v}$. The numerical simulations are performed for fixed domain of $25\lambda \times 25\lambda$. Bellin et al.[7] suggest that flow and transport solutions are not affected by refinements of the grid size involving more than four points per integral scale λ. Here, we choose the following grid sizes in the longitudinal and transverse directions, $\Delta x = \Delta y = \lambda/4$.

Steady-state subsurface flow in the 2-d domain is described by

$$\nabla . q(x) = \nabla .\left(-K(x)\nabla \phi(x)\right) = 0 \tag{10}$$

where $q(x)$ is the Darcy flux relative to the solid matrix and $\phi(x)$ is the hydraulic head. The flow equation is discretised by employing a mesh centered-finite difference scheme. The hydraulic heads at the nodal points are solved with the boundary conditions given above by a Gauss-Seidel method with successive over relaxation (SOR). In order to compute nodal heads, the interblock hydraulic conductivities are computed by geometric

averaging of adjacent conductivity values. This form of averaging ensures continuity of the head field and conservation of mass flux across block boundaries[17]. Darcy's law is applied to obtain the flux components $q(x)$ and nodal velocities are computed by finite differencing with a constant porosity value of $\eta = 0.3$.

In order to check the accuracy and efficiency of the finite difference approximation of the flow equation, we calculated the local mass balance error which can be defined[18] as

$$e = \left(Q_{out} - Q_{in}\right)/Q_{in} \times 100\% \tag{11}$$

where Q_{out} and Q_{in} are the total volume of flow out of the right boundary and into the left boundary, respectively. The mass balance error decreases when a smaller value of convergence criterion, μ, in the iterative scheme is used (the head changes in all nodes between two iterations is less than or equal to μ, iteration stops). We found that for each Monte Carlo simulation e is always less than 4% for each Hurst exponent H. These result, show that the finite difference solution is reasonably accurate, and accordingly, the obtained velocity field is sufficiently accurate to be used in the transport model. The mean velocity computed for Hurst exponent and with reference to 1500 Monte Carlo realisations was equal to unity leading to the conclusion that the mean velocity can indeed be assumed constant and unit in each realisation.

4 Stochastic Solute Transport Model

The random walk particle tracking method is commonly used in the field of statistical physics to model processes involving diffusion. This approach has also been used successfully to simulate reactive and non-reactive transport in the subsurface[4-7]. The basic idea is to approximate the spatial distribution of a transport quantity by a set of moving particles. The spatial location of particles are updated at each time step according to the following equation[23]

$$\mathbf{X}(t + \Delta t) = \mathbf{X}(t) + \left[\mathbf{V}(\mathbf{X}, t) + \nabla \cdot \mathbf{d}(\mathbf{V}(\mathbf{X}, t))\right]\Delta t + \left[2\mathbf{d}(\mathbf{V}(\mathbf{X}, t))\Delta t\right]^{1/2} \cdot \mathbf{Z} \tag{12}$$

where $\mathbf{X}(t + \Delta t)$ is the updated position of the particle that was at $\mathbf{X}(t)$ in the previous step, $\mathbf{V}(\mathbf{X}, t)$ is the velocity vector at the old position at time t, $\mathbf{d}$ is the local scale dispersion tensor, Δt is the time step, and $\mathbf{Z}$ is a vector of normally distributed random numbers of zero mean and unit variance. The second term on the right-hand side moves the particle advectively on the basis of the local velocity field at each point. The third term is important when stagnation regions or sharp fronts exist within the field[23]. The last term accounts for the local scale dispersion. The particle velocity needed in (12) is obtained by using a bilinear interpolation utilising the four velocity values surrounding the particle position. Eq. 12 is used to track 2000 non-interacting particles initially distributed along a strip of length 6λ normal to the mean flow direction of the domain. A constant dimensionless time step $\Delta t' = 0.05$ was chosen such that the fraction of the cell's length

traversed by a particle in a single time step is much smaller than one. The dimensionless time is given by $t' = \bar{v}t/\lambda$. The particle tracking experiments are performed in the inner core of the flow domain that is not affected by the boundaries. Particle tracking is terminated during each Monte Carlo run before the contaminant plume exits the inner core of the modelled domain.

5   Results from Flow and Transport Stochastic Analyses

Numerical generation of synthetic hydraulic conductivity fields, flow computations and transport simulations were carried out in a Monte Carlo manner for each Hurst exponent H. The simulation was terminated when the convergence criteria given in (9) were satisfied. For each Hurst exponent at least 1500 Monte Carlo runs were required to attain convergence. Numerical analyses were then carried out to calculate mean statistical quantities of interest from the flow and transport simulations.

Bellin et al.[7] show that the presence of boundaries influences the hydaulic head and velocity variabilities in two-dimensional domains. In Fig. 3, we plot results for the head variances σ_H^2 for different values of H averaged over 1500 Monte Carlo realisations. The head variability assumes a parabolic shape for each value of H. The parabolic shape of the curves reflects the statistical heterogeneity of the standard deviation in hydraulic head due to the constant head values specified on the boundaries of the flow domain. The head variance is zero at the fixed boundaries as expected and increases at the centre of the domain where it is maximum. The variance of hydraulic head σ_H^2, increases with H.
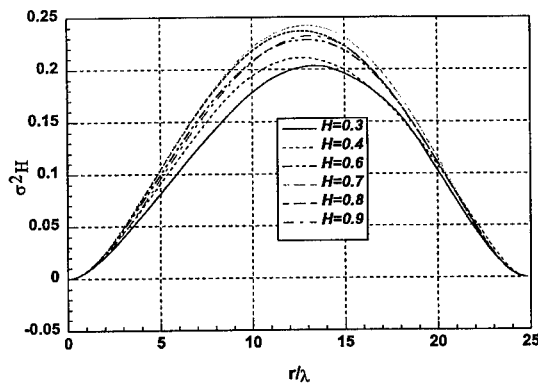

Figure 3: Head Variance σ_H^2 along the Longitudinal Axis for different values of H.

Figs. 4a and 4b show the average Monte Carlo longitudinal velocity variance σ_V^2 along the center line in the longitudinal direction (dashed curve), and also the center line along the transverse direction (solid curve), for $H = 0.4$ and 0.8 respectively.

(a)   (b)
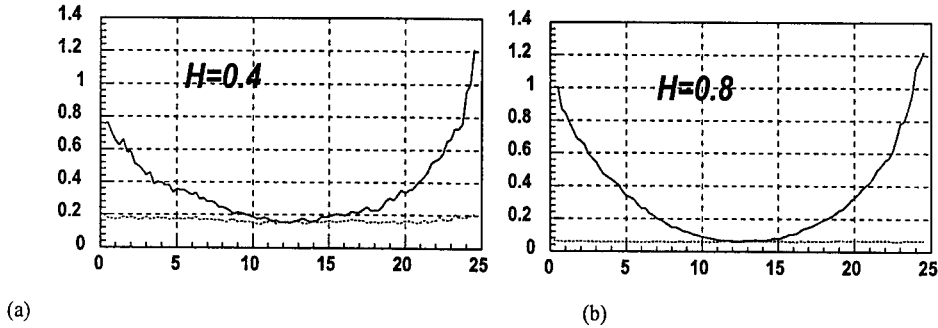
Figure 4: Velocity Variance σ_V^2 Along the Centreline in the Longitudinal (Dashed) and Transverse (Solid) Direction.

While in the longitudinal direction σ_V^2 is nearly constant, in the transverse direction it assumes a parabolic shape. The longitudinal velocity variance is maximum at the impervious boundaries and decreases in transverse direction with distance from the boundaries reaching the minimum value at the domain centre.

Travel time statistics are important[7] because they are robust in characterising the dispersion process blending all sources of uncertainty into a unique curve. Fig. 5 illustrates travel time statistical distributions (breakthrough curves) at three distinct absorbing barriers placed normal to the mean flow respectively at distances 5λ, 10λ, and 15λ computed by counting the number of particles the cross each barrier for each transport simulation time. The figure is the average of 1500 Monte Carlo simulations for the case $H = 0.7$. This result shows that as the plume travels further from the contaminant source, the increased spreading of the plume results in both a corresponding attenuation and reduction in peak value of the breakthrough curves.
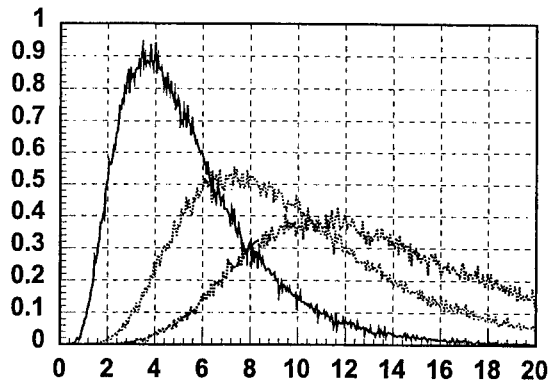
Figure 5:Travel Time Statistics for $H = 0.7$.

At the end of each Monte Carlo run, the moments of the dispersing plume are computed as follows: (a) the trajectory of the plume centroid:

$$R_i^m(t) = \frac{1}{N_p} \sum_{p=1}^{N_p} X_i^{m,p}(t),$$ (13)

where $X_i^{m,p}(t)$ ($i = 1,2$) are the coordinates of the p-th particle for the realisation m and N_p is the total number of particles; (b) the second-order central plume moment:

$$S_{ij}^m(t) = \frac{1}{N_p - 1} \sum_{p=1}^{N_p} \left[X_i^{m,p}(t) - R_i^m(t) \right]\left[X_j^{m,p}(t) - R_j^m(t) \right] \tag{14}$$

The moments (13) and (14) represent the overall plume behaviour in each one of the (equally probable) log conductivity fields. Average trajectories $\langle R_i(t) \rangle$, and the average of the second spatial moments $\langle S_{ij}(t) \rangle$ are obtained by taking expectations over MC independent Monte Carlo realisations:

$$\langle R_i(t) \rangle = \frac{1}{MC} \sum_{m=1}^{MC} R_i^m(t) \tag{15}$$

$$\langle S_{ij}(t) \rangle = \frac{1}{MC} \sum_{m=1}^{MC} S_{ij}^m(t). \tag{16}$$
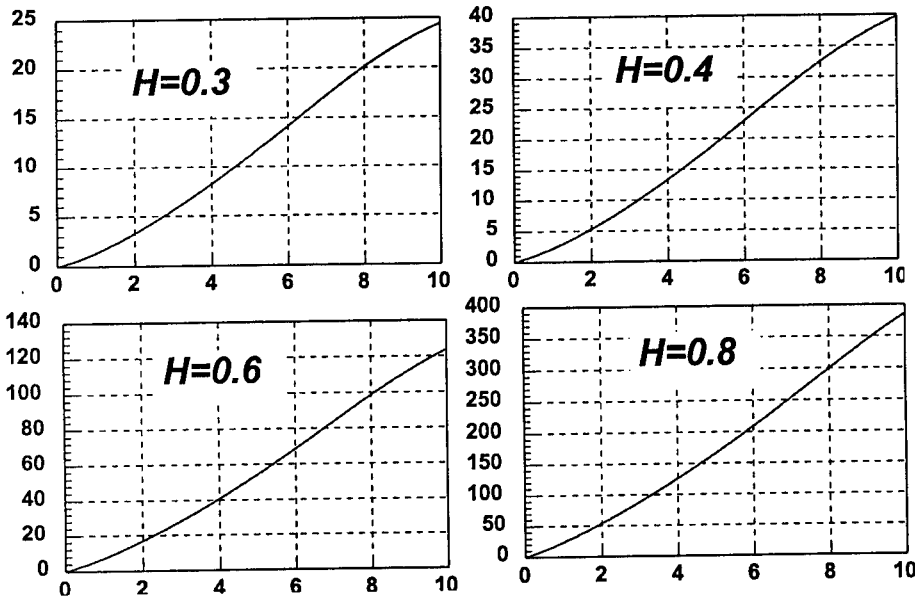
Figure 6: Particle Displacement Covariances $\langle S_{11}(t) \rangle - S_{11}(0)$ against $t' = \bar{v}t/\lambda$ for different values of H.

The expected values (15) and (16) represent the most probable estimate of the actual plume behaviour. Fig. 6 represents the average of 1500 simulated longitudinal second spatial moments $\langle S_{11}(t) \rangle - S_{11}(0)$ as a function of the dimensionless time $t' = \bar{v}t/\lambda$ for various values of H. Several observations can be made; as the value of the Hurst

exponent increases the second spatial moment increases with dimensionless time, a result that is consistent with the findings of Hassan *et al.*[5]. For each value of the Hurst exponent the second spatial moment increases non-linearly (non-Fickian) for all simulation times. This result can be explained by the fact that as the solute migrates in the porous medium, it continuously encounters spatially evolving scales of heterogeneity (fractal), hence the dispersion remains non-Fickian for all length scales below the scale of heterogeneity of the modelled domain (upper fractal cut-off scale R).

6   Summary

A two-dimensional numerical fractal-based Monte Carlo simulation model is developed to study flow and transport in heterogeneous media of spatially evolving heterogeneity. Realisations of the log fluctuating conductivity field are generated from a fractal or non-stationary distribution. The long range correlation of the hydraulic conductivity field is evident in the numerically obtained semivariograms. The flow simulations are carried out by solving a finite difference equation over a two-dimensional domain with uniform mean flow in one direction. The velocity field obtained from the conductivity realisations exhibits long range correlations. A random walk particle tracking technique is used to solve the solute transport problem for non-reactive solutes. The transport simulations indicate that higher Hurst exponent H (more persistence) of log conductivity results in more longitudinal spreading of the contaminant plume. Results also show that irrespective of the Hurst exponent the transport process is non-Fickian at all times for a dispersion process below the upper fractal cut-off scale R.

References

1. Gelhar, L. W. (1993). *Stochastic Subsurface Hydrology*. Prentice-Hall, Englewood Cliffs, New Jersey.
2. Rubin, Y. (1990). Stochastic analysis of macrodispersion in heterogeneous porous media. *Water Resources Research*, 26(1), 133-141.
3. Smith, L. and Freeze, R. A. (1979). Stochastic analysis of steady state groundwater flow in a bounded domain. 2. Two-dimensional simulations. *Water Resources Research*, 15(6), 1543-1559.
4. Smith, L., and Schwartz, F. W. (1980). Mass transport 1. A stochastic analysis of macroscopic dispersion. *Water Resources Research*, 16(2), 303-313.
5. Hassan, A. E., Cushman, J. H., and Delleur, J. W. (1997). Monte Carlo studies of flow and transport in fractal conductivity fields: Comparison with stochastic perturbation theory, *Water Resources Research*, 33(11), 2519-2534.
6. Bellin, A., Pannone, M., Fiori, A., and Rinaldo, A. (1996). On transport in porous formations characterised by heterogeneity of evolving scales. *Water Resources Research*, 32(12), 3485-3496.
7. Bellin, A., Salandin, P., and Rinaldo, A. (1992). Simulation of dispersion in heterogeneous porous formations: Statistics, first-order theories, convergence of computations. *Water Resources Research*, 28(9), 2211-2227.

8. Dagan, G. (1990). Transport in heterogeneous porous formations: Spatial moments, ergodicity, and effective dispersion. *Water Resources Research*, 26(6), 1281-1290.

9. Gelhar, L. W., and Axness, C. W. (1983). Three-dimensional stochastic analysis of macrodispersion in aquifers. *Water Resources Research*, 19(1), 161-180.

10. Sudicky, E. A. (1986). A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resources Research*, 22(13), 2069-2082.

11. Zhan, H., and Wheatcraft, S. W. (1996). Macrodispersivity tensor for non-reactive solute transport in isotropic and anisotropic fractal porous media: Analytical solutions. *Water Resources Research*, 32(12), 3461-3474.

12. Burrough, P. A. (1983). Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science*, 34, 577-597.

13. Molz, F. J., and Boman, G. K. (1995). Further evidence of fractal structure in hydraulic conductivity distributions. *Geophysical Research Letters*, 22(18), 2545-2548.

14. Kemblowski, M. W., and Wen, J. C. (1993). Contaminant spreading in stratified soils with fractal permeability distribution. *Water Resources Research*, 29(2), 419-425.

15. Neuman, S. P. (1990). Universal scaling of hydraulic conductivities and dispersivities in geologic media. *Water Resources Research*, 26(8) 1749-1758.

16. Mandelbot, B. B. (1983). *The Fractal Geometry of Nature*. W. H. Freeman, San Francisco.

17. Romeu, R.K., and Noetinger, B. (1995). Calculation of internodal transmisivities in finite difference models of flow in heterogeneous porous media. *Water Resources Research*, 31(4), 943-959.

18. Chin, D. A., and Wang, T. (1992). An investigation of the validity of first-order stochastic dispersion theories in isotropic porous media. *Water Resources Research*, 28(6), 1531-1542.

19. Mantoglou, A., and Wilson, J. L. (1982). The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resources Research*, 18(5), 1379-1394.

20. Yin, Z. M. (1996). New methods for simulation of fractional Brownian motion. *Journal of Computational Physics*, 127, 66-72.

21. Addison P.S. and Ndumu A.S. (1999), 'Engineering Applications of Fractional Brownian Motion: Self-Similar and Self-Affine Random Processes', *Fractals*, 7(2), 151-157.

22. Addison P.S. (1997), *Fractals and Chaos: An Illustrated Course*, Institute of Physics Publishing, Bristol.

23. Tompson, A. F. B., and Gelhar, L. W. (1990). Numerical simulation of solute transport in three-dimensional randomly heterogeneous porous media. *Water Resources Research*, 26(10), 2541-2562.

24. Voss, R. F. (1989). Random fractals: Self-affinity in noise, music, mountains, and clouds. *Physica D*, 38, 362-371.

RENDERING THROUGH ITERATED FUNCTION SYSTEMS

HUW JONES AND MAGNUS MOAR

Centre for Electronic Arts, Middlesex University, Cat Hill, Barnet EN4 8HT,
United Kingdom
E-mail: <d.h.jones, m.moar>@mdx.ac.uk

Iterated Function Systems (IFS) have been widely used for image compression and for generating fractal objects. Using barycentric coordinates and an extension of the concept, a range of rendering methods, such as Lambert, Gouraud and Phong shading, can be generated using IFS. By rendering a scattered collection of individual points using z-buffer and shadow buffer, the problems of clipping, hidden surface elimination and shadow generation are reduced to very simple forms. The method is relatively simple to program, but will not achieve the high speeds of current sophisticated rendering methods.

1   Iterated functions systems and the chaos game

Iterated function systems (IFS) have been used to generate models of fractal objects[1, 2] and for image compression[3, 4]. From the definitions of relatively few functions, often complex and unique figures can be drawn. Given n contractive transformations $\{f_0, f_1, f_2, \ldots f_{n-1}\}$ and an arbitrary starting point, a unique image in 2D or object in 3D is created by:

Given an initial point P_0
Loop for i = 1 to a large value
    Select function f at random from $\{f_0, f_1, \ldots f_{n-1}\}$
    Set point $P_i = f(P_{i-1})$
    Plot point P_i
End loop

This algorithm, often referred to as 'the chaos game', generates a Markov chain[5] of points, each new point dependent only on the previous one. Some stray points may be created initially (avoided by looping a few times before plotting or by setting P_0 inside the object), but points are soon 'attracted' into the image or object defined, the 'attractor' cannot then be escaped. Contractive transformations reduce the distance between distinct points. Uniform point density for each transformation is achieved when the probability of selection is proportional to contraction ratios (the fractional area or volume reduction), but uniform point density within the image is only achieved if the transformations have non-overlapping image sets.

The theoretical development of IFS[1, 2] and most subsequent uses concentrate on *affine* transformations, familiar to computer graphics experts as combinations of translations, scalings and rotations[6]. These standard rules have been relaxed in a number of studies, for example, Gröller[7] added 'tapering' and 'twisting' functions, Frame and Angers[8] used higher level polynomials, Jones and Campa[9] used randomised functions and Jones and Moar[10] used functions involving moduli. It has been shown by Hart[11] that, 'even the contractivity condition can be weakened to so called eventual contractivity'.

A particularly simple example of the 'chaos game' that generates a Sierpinski gasket or triangle (Fig. 1) is

Given three vertices V_0, V_1, V_2
Set point $P_0 = V_0$
Loop for $i = 1$ to a large value
  Select vertex V_i at random from $\{V_0, V_1, V_2\}$
  Set point $P_i = 0.5(V_i + P_{i-1})$
  Plot point P_i
End loop
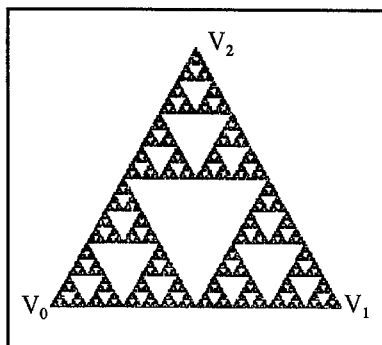
Figure 1. A Sierpinski gasket, based on an equilateral triangle, created by the chaos game algorithm. The method produces similar results for any form of triangle.

Point P_i is half way between the previous point P_{i-1} and the randomly selected vertex V_i. Setting $P_0 = V_0$ puts the initial point inside the attractor, so no stray points are created. As well as regular geometric forms, IFS can successfully create plant like images exhibiting self similarity[2, 10].

2   Barycentric coordinates

A point P in a triangle $V_0V_1V_2$ (Fig. 2) can be represented by barycentric coordinates (a, b) where

$$P = V_0 + a(V_1 - V_0) + b(V_2 - V_0),$$

using P, V_0, and so on as position vectors. This is a unique definition of P, with $a \geq 0$, b ≥ 0 and $(a + b) \leq 1$.
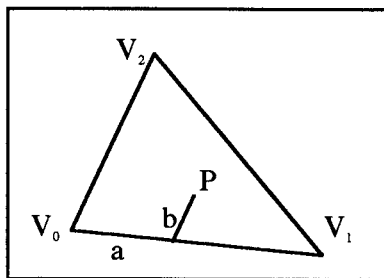
Figure 2. Barycentric coordinates (a, b) identify an interior point in a triangle.

These conditions are easily seen by identifying barycentric coordinates as a mapping of the general triangle $V_0V_1V_2$ to $V'_0V'_1V'_2$, where $V'_0 = (0, 0)$, $V'_1 = (1, 0)$ and $V'_2 = (0, 1)$, in a 2D Cartesian space (Fig. 3).
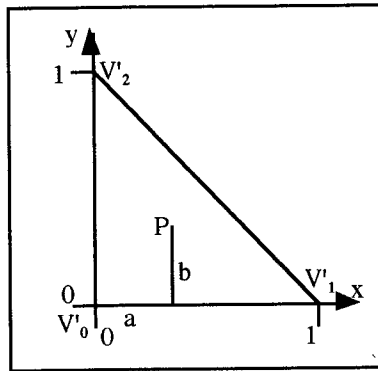
Figure 3. Mapping of the triangle $V_0V_1V_2$ into a Cartesian space to show conditions on a and b.

a and b are the fractions of edges V_0V_1 and V_0V_2 that are traversed in order to reach P from V_0. Rearranging the definition of P gives its position in terms of the triangle vertices as

$$P = (1 - a - b)V_0 + aV_1 + bV_2,$$

clearly a weighted mean of V_0, V_1, V_2 with weights $(1 - a - b)$, a and b. Conditions on a and b show these weights are non-negative and add to one, so P is the centre of mass of $(1 - a - b)$, a and b placed at V_0, V_1, V_2 respectively. This is analogous to use of blending functions in spline curves, is independent of order of selection of vertices and has implications for triangle shading.

3   The chaos game and barycentric coordinates

The functions implicit in the 'chaos game' map points from within the initial triangle to a half scaled triangle located at the randomly selected vertex, for example when vertex V_0 is selected, Fig. 4 shows the shaded half scale triangle, $V_0W_2W_1$ (W_i is the edge mid point opposite V_i). In the chaos game, if P_{i-1} has barycentric coordinates (a, b),

$$P_{i-1} = (1 - a - b)V_0 + aV_1 + bV_2,$$

the barycentric coordinates of the subsequent point $P_i = 0.5(V_i + P_{i-1})$ can be found for each selection of V_i.

1. When $V_i = V_0$,
    $P_i = (1 - 0.5a - 0.5b)V_0 + 0.5aV_1 + 0.5bV_2$,
    so the barycentric coordinates are (0.5a, 0.5b);
2. when $V_i = V_1$,
    $P_i = (0.5 - 0.5a - 0.5b)V_0 + 0.5(1 + a)V_1 + 0.5bV_2$
    $= (1 - 0.5(1 + a) - 0.5b)V_0 + 0.5(1 + a)V_1 + 0.5bV_2$,
    so the barycentric coordinates are (0.5(1 + a), 0.5b);

3. when $V_i = V_2$,
$$P_i = (0.5 - 0.5a - 0.5b)V_0 + 0.5aV_1 + 0.5(1 + b)V_2,$$
$$= (1 - 0.5a - 0.5(1 + b))V_0 + 0.5aV_1 + 0.5(1 + b)V_2,$$
so the barycentric coordinates are $(0.5a, 0.5(1 + b))$.

Thus, the version of the chaos game that generates a Sierpinski triangle can be rewritten as:

    Given three vertices V_0, V_1, V_2
    Set $a = 0$, $b = 0$
    Loop for $i = 1$ to a large value
      Select j at random from $\{0, 1, 2\}$
      Case
        $j = 0$: set $a = 0.5a$, $b = 0.5b$;
        $j = 1$: set $a = 0.5(1 + a)$, $b = 0.5b$;
        $j = 2$: set $a = 0.5a$, $b = 0.5(1 + b)$;
      End case
      Set point P_i to have barycentric coordinates (a, b)
      Plot point P_i
    End loop

This is slightly more complicated than the original form, but has interesting implications discussed later. Note that multiplication by 0.5 may be implemented very efficiently in most programming languages.

4   Rendering by IFS

4.1   *Triangle rendering*

Standard rendering algorithms 'fill' a polygon with colour. Some algorithms require that all the polygon faces are reduced to sets of triangles, so the triangle is the essential shape. The chaos game applied to triangle $V_0V_1V_2$ maps the original triangle to each of the sub-triangles $V_0W_2W_1$, $W_2V_1W_0$ and $W_1W_0V_2$ leaving internal triangle $W_0W_1W_2$ unfilled (Figs. 1, 5). The 'collage theorem'[2] indicates that a shape is generated by covering it with copies of itself, so to fill $V_0V_1V_2$, a fourth transformation is needed to map itself into the half scale congruent triangle $W_0W_1W_2$. The point P with barycentric coordinates (a, b) in $V_0V_1V_2$ should be mapped into P' with the same barycentric coordinates (a, b) in $W_0W_1W_2$ (Fig. 5). This sets

$$P' = (1 - a - b)W_0 + aW_1 + bW_2.$$

The W_i are the mid points of triangle edges, so

$$P' = 0.5(1 - a - b)(V_1 + V_2) + 0.5a(V_0 + V_2) + 0.5b(V_0 + V_1),$$

which can be rearranged as

$$P' = (1 - 0.5(1 - a) - 0.5(1 - b))V_0 + 0.5(1 - a)V_1 + 0.5(1 - b)V_2,$$

showing that the barycentric coordinates of P' in the original triangle $V_0V_1V_2$ are $(0.5(1 - a), 0.5(1 - b))$. This again finds the point P' as a weighted mean of triangle vertices V_0, V_1, V_2 with weights $0.5(a + b)$, $0.5(1 - a)$ and $0.5(1 - b)$ respectively. The weights are

all between 0 and 1 due to similar conditions on a and b, and they add to one, so this is a valid weighted mean. With equal probabilities of selection of non-overlapping transformations, each with equal contraction ratio of 0.5, triangle $V_0V_1V_2$ is filled with uniform point density by the following adapted chaos game.

```
Given three vertices V₀, V₁, V₂
Set a = 0, b = 0                    // arbitrarily chooses V₀ as the first point
Loop for i = 1 to a large value
  Select j at random from {0, 1, 2, 3}
  Case
  j = 0: set a = 0.5a, b = 0.5b;
  j = 1: set a = 0.5(1 + a), b = 0.5b;
  j = 2: set a = 0.5a, b = 0.5(1 + b);
  j = 3: set a = 0.5(1 – a), b = 0.5(1 – b);
  End case
  Set point Pᵢ to have barycentric coordinates (a, b)
  Plot point Pᵢ
End loop
```
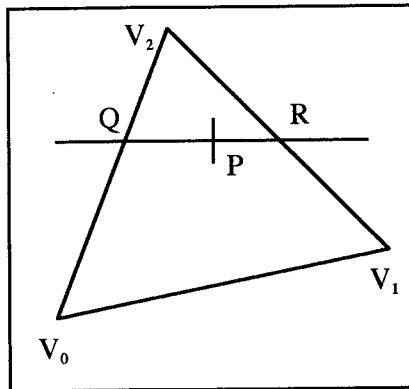

Figure 4. Scan line rendering: the internal point P lies on a scan line through Q and R.

In scan line rendering, the colours of interior points of polygons are calculated by linear interpolation, initially along edges and then along a scan line, of pre-calculated colour (for Gouraud shading) or surface normals (for Phong shading) at polygon vertices[6]. In Fig. 4, a scan line cuts the 'active' edges V_2V_0 and V_2V_1 at Q and R. If $I(V_i)$ is the colour intensity at Vi, the colour at Q is

$$I(Q) = (1 – q)I(V_2) + qI(V_0),$$

where $q = V_2Q/V_2V_0$. With $r = V_2R/V_2V_1$

$$I(R) = (1 – r)I(V_2) + rI(V_1),$$

and with $p = QP/QR$,

$$I(P) = (1 – p)I(Q) + pI(R).$$

In practice, these formulae are used to pre-calculate offset colour changes and positional changes of P, Q and R for horizontal and vertical pixel moves.

The method has a slight distorting effect, as the perspective projection of the 3D triangle is non-affine. Points closer to the observer should display smaller colour change per pixel, but the method gives a uniform change in the image space, rather than in the world space[6]. The method is, in any case, a distortion of the correct physical model. The algorithms for Gouraud and Phong shading allow updating of the active edge list when the scan line passes a vertex, so 'spans' like QR can be easily identified and updated for polygons other than triangles.

For a triangle with an interior point P known through its barycentric coordinates (a, b), colour can be established directly from the weighted average of the colour (or surface normal) values at the vertices, without reference to scan line directions. For Gouraud shading,

$$I(P) = (1 - a - b)I(V_0) + aI(V_1) + bI(V_2).$$

For Phong shading, the mean surface normal vector is

$$n(P) = Normalise[(1 - a - b)n(V_0) + an(V_1) + bn(V_2)],$$

where $n(P)$ is the surface normal allocated to point P, and Normalise is a function to reduce a vector to unit length. The argument of a 'Normalise' function cannot have zero length, this would only occur in the case of a degenerate or non-manifold object.

This gives an easy way to fill triangles using IFS with barycentric coordinates, which can be shown to give exactly the same results as standard Gouraud and Phong shading. The barycentric method is independent of the scan line direction, and this equivalence indicates similar independence of the traditional methods from the scan line direction for *triangle* shading, which is not the case for polygons with more edges.
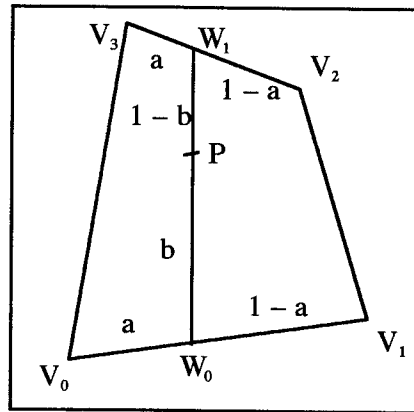
Figure 5. Extended barycentric coordinates identify an inerior point P in a quadrilateral.

4.2 Quadrilateral rendering

We can identify an internal point P of a quadrilateral $V_0V_1V_2V_3$ using 'extended barycentric coordinates' (a, b). There is a unique line W_0W_1 which passes through P so that the edges V_0V_1 and V_3V_2 are divided in equal proportion, $a:(1 - a)$ (Fig. 5). In other words,

$$W_0 = (1 - a)V_0 + aV_1,$$

$$W_1 = (1 - a)V_3 + aV_2,$$

where $0 \le a \le 1$. If P lies a proportion b along W_0W_1

$$P = (1 - b)W_0 + bW_1,$$

where $0 \le b \le 1$. This effectively maps any interior point of the quadrilateral into a unit square (Fig. 6), where

$V'_0 = (0, 0)$, $V'_1 = (1, 0)$, $V'_2 = (1, 1)$ and $V'_3 = (0, 1)$.
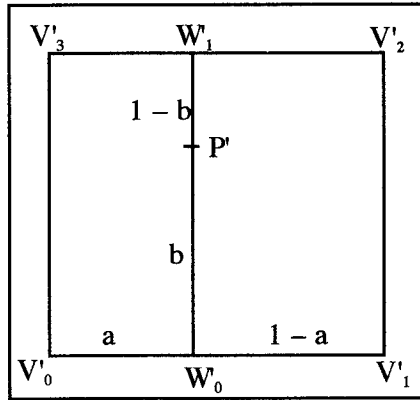
Figure 6. Mapping of the quadrilateral of figure 5 into a unit square

This is equivalent to the effect of Fig. 3 for a triangle. The meaning of an internal point is obvious for a planar convex quadrilateral (Fig. 5). For non-planar vertices, the point P lies on a bi-linear patch, generating a form of hyperbolic paraboloid[12]. For planar concave quadrilaterals, P may lie outside the triangle boundary, but within the convex hull of the vertices. Such quadrilaterals can be identified and split into triangles when rendering, although this allows the ambiguity of two possible interpretations.

Substituting for W_0 and W_1 into the expression for p gives

$$P = (1 - b)\{(1 - a)V_0 + aV_1\} + b\{(1 - a)V_3 + aV_2\},$$

$$P = (1 - a)(1 - b)V_0 + a(1 - b)V_1 + abV_2 + (1 - a)bV_3,$$

giving P as a weighted mean of the four vertices. This is valid as each of the four weights $(1 - a)(1 - b)$, $a(1 - b)$, $(1 - a)b$ and ab lies between zero and 1 from conditions on a and b, and it is easy to establish that their sum is one. This gives a direct method for proportional shading from vertex colours or normals, which is the same as standard scan line methods only when W_0W_1 happens to be aligned with a scan line (Fig. 7). The IFS method is an improvement on standard methods, in that the shading allocated to a point is not dependent on quadrilateral orientation.

The chaos game applied to the four vertices of a square (Fig. 6) will fill the square[13]. By selecting V'_i to be one of the vertices V'_0, V'_1, V'_2 or V'_3 at random, and finding the mid point of P' and V'_i, we have the essential step of the chaos game that will fill the square $V'_0V'_1V'_2V'_3$. It should be clear that the required mid points have coordinates

- P' and V'_0: (0.5a, 0.5b);
- P' and V'_1: (0.5(1 + a), 0.5b);
- P' and V'_2: (0.5(1 + a), 0.5(1 + b));
- P' and V'_3: (0.5a, 0.5(1 + b)).

Applying the usual chaos game directly to the vertices of a general quadrilateral $V_0V_1V_2V_3$ (Fig. 5) will only fill it if it is a parallelogram. When the quadrilateral is skew, a fractal 'Sierpinski tetrahedron' will result[9]. If the chaos game is applied to the extended barycentric coordinates of a general quadrilateral according to the algorithm specified below, it fills all 'interior' points as explained earlier in this section, not now with uniform density, but still giving weights that lead to easy rendering from vertex colours or normals.

```
Set (a, b) to (0, 0)
Loop for i = 1 to a large value
  Select j at random from {0, 1, 2, 3}
  Case
    j = 0: set a = 0.5a, b = 0.5b;
    j = 1: set a = 0.5(1 + a), b = 0.5b;
    j = 2: set a = 0.5(1 + a), b = 0.5(1 + b);
    j = 3: set a = 0.5a, b = 0.5(1 + b);
  End case
  Set P_i to have extended barycentric coordinates (a, b)
  Plot point P_i
End loop
```

For Gouraud type shading, we have

$$I(P) = (1 - a)(1 - b)I(V_0) + a(1 - b)I(V_1) + ab\,I(V_2) + (1 - a)bI(V_3),$$

where $I(V_i)$ is the colour allocated to vertex V_i, as above. For Phong type shading, normals are interpolated by

$$n(P) = \text{Normalise}[(1 - a)(1 - b)n(V_0) + a(1 - b)n(V_1) + \\ abn(V_2) + (1 - a)bn(V_3)].$$

Shading depends only on vertex values, so is independent of scan line direction, unlike standard methods. Symmetry of expressions shows that the order of selection of vertices does not affect shading.

5 An example: Gouraud shading of a cone

We consider Gouraud shading of a cone as a simple example. If the cone is represented polyhedrally as a pyramid with n sloping triangular faces and a regular n-gon as its base (Fig. 7), there is a problem of how to allocate surface normals to the triangle vertices which meet at V. If we average normals for all triangles meeting at this point, the applied normal points directly out of the vertex along the cone's axis, giving the vertex an incorrect flattened shading. Using the triangle's own normal at V for each triangle avoids this, but this shows creasing along triangle edges near V, as there is no continuity across triangle edges. The singular point at V has different normals, hence shading values, close to V.
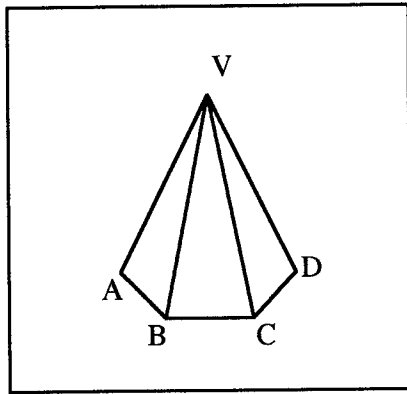
Figure 7. Polygonal approximation of a cone as a pyramid on a regular polyhedral base.
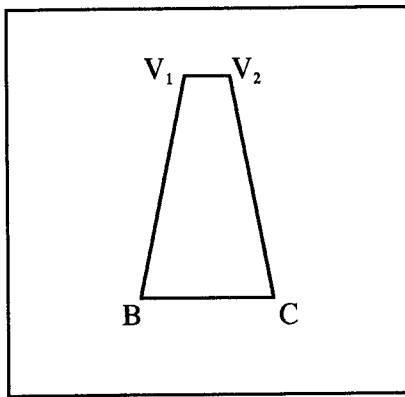
Figure 8. Representation of triangle BCV as a degenerate quadrilateral.

If vertex triangles are replaced by quadrilaterals with a zero length edge at V (Fig. 8), the problem can be overcome. At V_1, the surface normal is taken as the normalised mean of normals of faces VAB and VBC and at V_2 the normal is the normalised mean of the normals of faces VBC and VCD. If V_1 and V_2 are made coincident at V, this gives continuity at V, avoiding creasing. The now degenerate trapezium V_1BCV_2 has the same shape as the isosceles triangle VBC. There is still a singular point at V, as more than one normal is allocated to the same point, with a multiplicity of colour values possible at the vertex on rendering, but this is anti-aliased by the random nature of the IFS. Figure 9 clearly shows creasing from the standard method and its elimination by the degenerate quadrilateral method, both rendered using an IFS on an eight-sided based pyramid. (This image was rendered on a 16 bit machine, so some colour banding is seen).
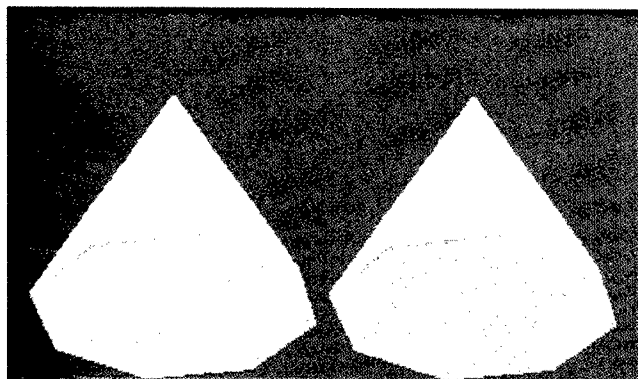
Figure 9. Gouraud shaded cones represented as octagonal based pyramids. The left image shows creasing at the vertex, the right image, which uses degenerate quadrilaterals, has a smoother representation.

6 Summary

A method for using IFS to render non-fractal forms in a way similar to standard scan line algorithms has been shown. Compared to scan line methods, IFS rendering uses different algorithms for triangles and quadrilaterals, not dealing with higher orders. The algorithm selects world points at random, so stochastic anti-aliasing occurs by painting each pixel with a sample of its possible colours. This is an advantage for still images, but may give undesirable 'boiling' on animation. Standard methods visit each pixel once per polygon, the IFS method has redundancy, with pixels being visited many times without a guarantee that all are filled. IFS methods are unlikely to challenge established methods on speed. However, programming them is relatively simple, with clipping, hidden surface and shadow problems reduced to simple decisions.

An example using Gouraud type shading on a basic shape is shown (Fig. 9). There is much scope for extension. Work on Phong type shading with specular highlights and on texture mapping, for which the method should be particularly suited, is underway at

Figure 10. IFS images showing fractal and non-fractal objects using z-buffer and shadow buffer and showing the potential of the method for texture mapping.

the time of writing. As an indication of the scope of the method, two images incorporating fractal trees (with non-affine transformations), shadows and texturing (for the trellis and mown grass effect)[10] are shown in Fig. 10.

7 Acknowledgments

References

1. Barnsley M. F., Fractal Modelling of Real World Images. In *The Science of Fractal Images*, ed. by Peitgen H. O. and Saupe D. (Springer-Verlag, New York, 1988) pp. 219–242.
2. Barnsley M. F. *Fractals Everywhere (2 ed)* (Academic Press, San Diego, 1992).
3. Fisher Y. *Fractal Encoding – Theory and Applications to Digital Images* (Springer-Verlag, New York, 1994).
4. Lu N. *Fractal Imaging* (Academic Press, San Diego, 1997).
5. Cox D. R. and Miller H. D. *The Theory of Stochastic Processes* (Methuen, London, 1965).
6. Foley J. D., van Dam A., Feiner S. K. and Hughes J. F. *Computer Graphics, Principles and Practice (2 ed)* (Addison-Wesley, Reading, 1990).
7. Gröller E. Modeling and Rendering of Nonlinear Iterated Function Systems, *Comput. & Graphics*, **18**(5), (1994) pp. 739-748.
8. Frame M. and Angers M. Some Nonlinear Iterated Function Systems. *Comput. & Graphics*, **18**(1), (1994) pp. 119-125.
9. Jones H. and Campa A. Abstract and natural forms from iterated function systems. In *Communicating with Virtual Worlds*, ed. by Magnenat Thalmann N. and Thalmann D. (Springer-Verlag, Tokyo, 1993) pp. 332-344.
10. Jones H. and Moar M. Iterated Function Systems and Non Affine Transformations: Examples in 2D and 3D, *15th Spring Conference on Computer Graphics*, (University of Bratislava, Budmerice, Slovakia, 28 April - 1 May, 1999) pp. 201-209.
11. Hart J. C. Iterated Function Systems and Recurrent Iterated Function Systems. In Fractal Models for Image Synthesis, Compression and Analysis, *ACM SIGGRAPH 1996, Course 27*, (ACM, New Orleans, 6 August 1996).
12. Cundy H. M. and Rollett A. P. *Mathematical Models (2 ed)* (Oxford University Press, London, 1961).
13. Jones H. Dürer, Gaskets and the Chaos Game, *Computer Graphics Forum*, **9**(3) (1991) pp. 327-332.

THE σ-HULL – THE HULL WHERE FRACTALS LIVE CALCULATING A HULL BOUNDED BY LOG SPIRALS TO SOLVE THE INVERSE IFS-PROBLEM BY THE DETECTED ORBITS

ERWIN HOCEVAR

Technical University of Vienna, Institute of Automation 183/2
A-1040 Treitelstr. 3, Vienna, Austria
E-mail: hocevar@zdvaxa.arcs.ac.at

Global IFS seem to be suited best for compressed encoding of natural objects which are in most cases self affine even if not always exactly. Since affine transformations - the IFS-Codes - resp. the union of all their orbits generate an object (an IFS-Attractor), the detection of a non minimal set of these orbits solves the inverse IFS-Problem by calculating a superset of IFS-Codes, which has to be minimized.
Here a method is presented how these orbits (in particular those on the object boundary) can be calculated. Therefore a generalized convex hull - the σ-Hull - is defined. This fractal hull is bounded by log spirals, that curves formed by the orbits. It is shown that log spirals can be represented by a continuos function of powers of affine maps and that by using this "spiral equivalent" the generating transformations of the orbits by which an IFS-object is bound, can be calculated in the x/y-plane.
Further more this representation can be used for the classification of the detected orbits, necessary to calculate the IFS-Codes of a minimal IFS from their generating transformations, subsequently.

1   Introduction

1.1   Problem Specification - Objects to be encoded

The idea is to globally encode[1] objects of an image appearing in the nature and therefore often self similar to a high degree but difficult to encode or to compress by a mathematical formula. This *global IFS-Approach*[2] - representing an object by the union of affine contractive transformed copies (subobjects generated by the IFS-Codes) of the object itself - seems to be used best to solve the inverse (global) IFS-Problem.

Thus the *Collage Theorem* which can be used to generate whole objects of an image is given: A set $\{\omega_i: R^2 \to R^2 \mid \omega_i$ affine, contraction$\}$ of complete Euclidean space (R^2, d) is an *Iterated Function Systems (IFS)* with contractivity factor s, for $i \in N$, the distance d, $0 \le s < 1$ and $s \cdot h(B, C) \ge h(W(B), W(C))$, if for the Hutchinson operator W^a holds

$$W(A) = \bigcup_{\omega_i \in W} \bigcup_{x \in A} \omega_i(x) \text{ and } A = W(A).$$

The compact set A is called an *attractor* of W.
The *Collage Theorem* (CT) for attractors close to given compact sets:

$$h(L, W(L)) \le \varepsilon \Rightarrow h(L, A) \le {}^{\varepsilon}\!/_{(1-s)}$$

where h is called the *Hausdorff Distance*

---

[a] W: $\mathcal{H}(R^2) \to \mathcal{H}(R^2)$ with W(A) = {W(a) | a∈A}, A∈$\mathcal{H}(R^2)$ is a contractive map of ($\mathcal{H}(R^2)$,h) the space of nonempty compact subsets of R^2 where h(C,B) = max{d(C,B), d(C,A)} and C, B∈$\mathcal{H}(R^2)$.

$L \subseteq R^2$ is a given compact set of (R^2,d) and $\epsilon \geq 0$. For $\epsilon = 0$ $(L = A)$ the CT for exactly self affine attractors is obtained. In principle the CT can be applied to any object if sufficient many IFS-Codes ω_i are used. However, to compress an object the number of parts ω_i (L) (the subobjects) have to be small for a small ϵ.

For the discrete pixel space the *discrete CT*[3] is defined as a set $\{\omega_m: P^2 \to P^2 \mid \omega_m$ affine, contraction$\}$ of the pixel space (P^2,d) - $P = \{0,...,R-1\} \subset N$ - is a *Discrete Iterated Function Systems (DIFS)* with contractivity factor s, for $m \in P$, R^2 the number of image screen pixels, the block distance d, $0 \leq s < 1$ and $s \cdot h(B,C) \geq h(W(B), W(C))$, if for the Hutchinson operator $W: \mathcal{H}(P^2) \to \mathcal{H}(P^2)$ holds

$$W(A) = \bigcup_{m \in P} \bigcup_{x \in A} \omega_m(x) \text{ and } A = W(A).$$

The set A is called an discrete *attractor* of W. The discrete attractor $A \in \mathcal{H}(P^2)$ is not unique, but there always exist a unique maximal attractor $A_{max} \in \mathcal{H}(P^2)$ with $A = W(A) \Rightarrow A \subseteq A_{max}$, which contains all discrete attractors.

The *Chaos Game* generates an attractor A by the iterated random application of the different ω_m to one image point i.e., by the union of orbits formed by all combinations of ω_m - the basic principle of the presented solution.

In contrast to global IFS, an *encoding* based on *Partitioned IFS* (PIFS)[4] divides the image into larger and smaller segments (polygons). The smaller target segments are interpreted as affine contractive copies of the larger ones and then the CT is applied. Since only IFS-Properties of affine relationships constrained to the segment form can be exploited, all of the many image parts are only approximated by an IFS-encoding.

1.2   The Solution - How to encode global IFS-Objects

Calculating boundaries of an object

In general the boundary of an object contains the most information how an object is built. Points inside an object normally cannot be related, because they are difficult to distinguish. The common method to determine the boundary of an object - especially if it is defined in a discrete (pixel) space - is to form the hull of an object.

We first can try it by the convex hull $\mathcal{H}(A)$, which is defined as "Intersection of half planes containing a point set (the attractor) A". It soon appears that a hull of such a kind is not usable for IFS-Attractors, because only the boundary of convex objects - what an IFS certainly is not - can be determined.

Second we can try a more general hull - the α-Hull[5] $\mathcal{H}_\alpha$ - defined as follows :
The α-Hull $\mathcal{H}_\alpha(A)$ (see *Fig. 1*) is defined as the intersection of all generalized disks B(c, r) resp. B(c, -r) of non negative radius $r = 1/\alpha$ and the centre c containing the attractor A, where B(c, -r) is the complement of the disk B(c,r).[6]

The boundary of an α-Hull is formed by α-Extremes (see *Fig. 1*):
A point $s \in A$ is termed *α-Extreme* (corner of $\mathcal{H}_\alpha(A)$) in A, if there exists a B(c, $1/\alpha$) such that s lies on its boundary and contains all other points of A.[6]

Now it is possible to analyse concave objects and even objects concave with different curvature using different radii of the generalized disks. But we need a relationship of boundary points (extremes) which provides more information about the object. Using the

α-Hull we get only a relationship of points laying on the same edge of a disk of different radius forming the α-Hull. But the boundary points of an IFS-Attractor do not lay on circles. They form orbits[b] in the discrete space which are given as subset of W^+ [c] for an IFS W (as it is easily to observe in the application of the Chaos Game). These orbits lie on logarithmic spirals[1], what will be shown in this paper. So the best suited hull to determine an IFS-Attractor seems to be the σ-Hull which is bounded by log spirals and will be detailly defined and described in the main part of this paper.

Calculating a minimal IFS

Once found the orbits on the boundaries of different σ-Hulls they have to be classified to calculate the IFS-Codes:

> Classes C_i = {orbits which can be affinely mapped onto eachother} are formed.
> The potential IFS-Codes ϖ_i are the generating transformations of the largest orbits[d] of each class.

Then the final IFS-Codes ω_i are the generating transformations of the largest orbits within each C_i which can be mapped onto eachother by (have an inter orbit transformation in) $\{\omega_i\}^+$.

Thus the final and also minimal IFS can be calculated by the following iterative process

(a) which restrains the classes to those orbits which can be mapped onto each other only by ϖ_i detected so far
> ⇒ more and more classes are generated (and the set $\{\varpi_i\}$ is increased)

(b) which enlarges the classes to those orbits which can now be mapped onto each other only by ϖ_i detected so far
> ⇒ fewer and fewer classes are generated (and the set $\{\varpi_i\}$ is decreased)

This process is repeated until no more (in (a)) resp. no fewer (in (b)) classes are generated. Then the generating transformations of the largest orbits of each remaining class form a minimal set of the IFS-Codes of the object.

Since a σ-Hull is only useful to find the orbits on the boundary of an object, orbits of inner subobjects (having no boundary in common with the object) cannot be found in the first step. But if you form the difference between the object A_0 and the point set generated by the IFS-Codes in the first step one will get a difference object A_1 in the second step. Those affine transformations mapping boundary orbits (detected by different σ-Hulls) of the object A_0 to those of the difference object A_1 are the IFS-Codes for the inner subobjects now lying outside. The second step has to be repeated for the object A_i ($i \in \mathbb{N}$) as long as the difference object A_{i+1} is not empty. Adjusting the obtained results and eliminating possible deviations by comparing the inverse IFS-mapping of the hull of inner subobjects with the object hull, also the inverse IFS-Problem for overlapping subobjects can be solved.

---

[b] An *orbit O* is a point sequence $o_n = \phi^{(n)}(x)$ $\forall n \in \mathbb{N}$ where $\phi^{(n)} = \prod_n \phi$, ϕ affine and $x \in \mathbb{R}^2$.

The point sequence (orbit) $\phi^{(n)}(p)$ converges for a contractive affine maps ϕ to the *fixed point* f of ϕ

[c] $W^* = \{ \prod_{n \in \mathbb{N}} \omega_n \mid \omega_n \in W \cup \{\varepsilon\} \}$ and $W^+ = W^* \setminus \{\varepsilon\}$ with $W = \{\varpi_i \mid \varpi_i$ affine$\}$ and ε the *Identity Map*.

[d] The *largest orbits* of a subobject $\rho(A)$ ($\rho \in W \cup \{\varepsilon\}$) are those orbits $O = \{o_0, ..., o_n\}$ with a minimal contraction, where $\phi^{(n)}(o_0) \in A$ for $n \geq 0$ and $\phi^{(n)}(o_0) \notin A$ for $n < 0$ and $\phi \in W^+$.

2 Definition and Application of the σ-Hull

2.1 The σ-Hull (fractal hull) in $\mathbb{R} \times \varphi$
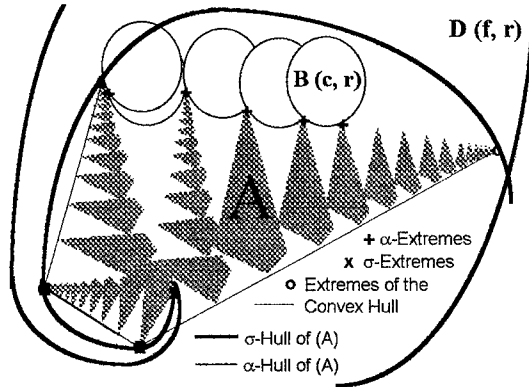
Figure 1. The convex Hull, the α- and σ-Hull of an Object

We first introduce the notion of generalized spiral disks. Let $D^*(f, r)$ denote a spiral disk defined by the point set $\bigcup_i^\infty p_i$ with fixed point (centre) f, the non negative radius $r = Re^{k\varphi}$ where $p_i \in \mathbb{R} \times \varphi$ and $d(f, p_i) \le r$ and $0 \le \varphi \le 2m\pi$, $m \in \mathbb{N}$.

A *generalized spiral disk*

$$D'(f, r) \equiv \begin{cases} D^*(f, -r)^C & r < 0 \\ D^*(f, r) & r \ge 0 \end{cases} \tag{1}$$

where X^C is the complement of X.

Since $D'(f, r)$ is not solely bounded by a log spiral (but also by a part of the $\varphi = 0$ axis) a generalized spiral disc segment has to be constructed bounded only by equally curved log spirals having the same fixed point.

Thus a *generalized spiral disk segment* is given by *(Fig. 1)*:

$$D(f, r) = D'(f, Re^{k\varphi}) \cap D'(f, -Re^{k(\pi/4+\varphi)}) \cap D'(f, Re^{k(\pi/2-\varphi)}) \tag{2}$$

Note: For negative k one will get a spiral disc segment reflected by the $\varphi = 0$ axis !

$C_\sigma(S)$ is the *set of fixed points* of spiral disc segments of radius $1/_\sigma$ that have S as subset

$$C_\sigma(S) = \{x \in \mathbb{R} \times \varphi \mid S \subseteq \overline{D(x, 1/\sigma)}\} \tag{3}$$

where $\overline{X}$ is the set closure of X.

Let X be an arbitrary (finite or connected set) in $\mathbb{R} \times \varphi$ then the *intersection of all generalized spiral disc segments* with varying fixed point $x \in X$ and a fixed radius $1/_\sigma$

$$\text{is denoted by } M_\sigma(X) \equiv \begin{cases} R x \varphi & X = \varnothing \\ \bigcap_{x \in X} D(x, {}^1/\sigma) & X \neq \varnothing \end{cases} \tag{4}$$

where $\varnothing$ is the empty set.

The σ-*Hull* of S is the intersection of all generalized spiral disk segments of radius $^1/_\sigma$ that contain all points of S. Thus the σ-*Hull (fractal hull)* is given by

$$\mathscr{H}_\sigma(S) \equiv M_\sigma(C_\sigma(S)) \tag{5}$$

A point $s \in S$ is termed σ-*Extreme* (corner of $\mathscr{H}_\sigma(S)$) in S, if there exists a $D(c, {}^1/_\sigma)$ such that s lies on its boundary and contains all other points of S (see *Fig. 1*).

Thus the set of σ-*Extremes* is defined by

$$E_\sigma(S) \equiv \{s \in S \mid \exists c \in C_\sigma(S) : s \in \partial(c, {}^1/_\sigma)\} \tag{6}$$

where ∂X is the boundary of S.

As in the limit case of infinite radius (i.e. $^1/_k$ close to 0) resp. of radius $= R$ ($k = 0$) the generalized spiral disc segment becomes a half plane resp. a circle of radius R, this definition includes the definition of the convex Hull $\mathscr{H}(S)$ resp. of the α-Hull $\mathscr{H}_\alpha(S)$ as well.

2.2 Defining the σ-Hull in $\mathbb{R}^2$

To represent a log spiral $r = Re^{k\varphi}$ defined in the polar plane ($\mathbb{R} \times \varphi$) an equivalent for the x/y-plane ($\mathbb{R} \times \mathbb{R}$) has to be found. The best suited equivalent seems to be one formulated in terms of affine transformations. In addition, this representation will relate log spirals and orbits (lying on the log spirals) which union forms an IFS and where the generating transformations of the largest of these orbits are the IFS-Codes.

Theorem 1 - *Log spiral equivalent* - A log spiral $r = Re^{k\varphi}$ can be represented by a continuos function of powers of an affine transformation ω.

$$Re^{k\varphi} \equiv \omega^n = R_0(\alpha) \, S(c_x^{\,n}, c_y^{\,n}) \, R_0(-\alpha) \, R_0(n(-\beta)),$$

where $n \in \mathbb{R}$, $\alpha \in [-\pi, \pi)$, $\beta \in [-\pi, \pi)$ and $\beta \neq 0$

Lemma 1 Each affine transformation can be represented by a symmetric matrix and a rotation matrix.

$$\forall \omega = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2 \times 2} \, \exists \beta \in [-\pi, \pi) \, \exists \psi = \begin{pmatrix} r & s \\ s & t \end{pmatrix} \in \mathbb{R}^{2 \times 2} :$$

$$\omega = \psi R_0(-\beta) = \begin{pmatrix} r\cos\beta + s\sin\beta & s\cos\beta - r\sin\beta \\ s\cos\beta + t\sin\beta & t\cos\beta - s\sin\beta \end{pmatrix} \text{ with } \tan\beta = \frac{b-c}{a+d}$$

$$r = \frac{a^2 + b^2 + ad - bc}{\sqrt{(a+d)^2 + (c-b)^2}}, \quad t = \frac{c^2 + d^2 + ad - bc}{\sqrt{(a+d)^2 + (c-b)^2}} \quad \text{and} \quad s = \frac{ac - bd}{\sqrt{(a+d)^2 + (c-b)^2}}$$

Proof of Lemma 1: in [3]

Lemma 2 Each symmetric matrix can be represented by a rotation α, a scaling matrix and a rotation $-\alpha$.

$$\forall \psi = \begin{pmatrix} r & s \\ s & t \end{pmatrix} \in \mathbb{R}^{2\times2} \; \exists \; \alpha \in [-\pi, \pi) \; \exists \; c_x, c_y \in \mathbb{R} : \psi = R_0(\alpha) \; S(c_x, c_y) \; R_0(-\alpha)$$

with the eigenvalues $c_x = \dfrac{r+t+\sqrt{(r-t)^2+4s^2}}{2}$, $\qquad c_y = \dfrac{r+t-\sqrt{(r-t)^2+4s^2}}{2}$

and $\alpha = \arctan\dfrac{c_x-r}{s}$ the angle of the eigenvector $\vec{e}_x = \begin{pmatrix} c_x-r \\ s \end{pmatrix}$

Proof of Lemma 2: in [3]

Lemma 3 Each affine transformation ω maps a point $\overline{x} \in Re^{k\varphi}$ again to a point $\overline{x}_1 \in Re^{k\varphi}$.

$$\forall \omega \in \mathbb{R}^{2\times2} \text{ and } \overline{x} \in Re^{k\varphi} : \omega\overline{x} = \overline{x}_1 \text{ and } \overline{x}_1 \in Re^{k\varphi}$$

Proof of Lemma 3:

$$r = Re^{k\varphi} \text{ resp. } \sqrt{x^2+y^2} = Re^{k\arctan\frac{y}{x}}$$

$$x = r\cos\varphi = Re^{k\varphi}\cos\varphi \text{ and } y = r\sin\varphi = Re^{k\varphi}\sin\varphi$$

$$\omega\overline{x} = \omega\begin{pmatrix} x \\ y \end{pmatrix} = \omega\begin{pmatrix} Re^{k\varphi}\cos\varphi \\ Re^{k\varphi}\sin\varphi \end{pmatrix} = Re^{k\varphi}\begin{pmatrix} a\cos\varphi+b\sin\varphi \\ c\cos\varphi+d\sin\varphi \end{pmatrix} = \overline{x}_1$$

$$\varphi = \arctan\frac{y}{x} = \arctan\frac{a\cos\varphi+b\sin\varphi}{c\cos\varphi+d\sin\varphi} \text{ and}$$

$$r = \sqrt{x^2+y^2} = Re^{k\varphi}\sqrt{(a\cos\varphi+b\sin\varphi)^2+(c\cos\varphi+d\sin\varphi)^2}$$

$$Re^{k\varphi}\sqrt{(a\cos\varphi+b\sin\varphi)^2+(c\cos\varphi+d\sin\varphi)^2} = Re^{k\varphi}e^{k\arctan\frac{c\cos\varphi+d\sin\varphi}{a\cos\varphi+b\sin\varphi}} \Big| :e^{k\varphi}$$

$$\underbrace{\sqrt{R^2(a\cos\varphi+b\sin\varphi)^2}}_{x^2}+\underbrace{R^2(c\cos\varphi+d\sin\varphi)^2}_{y^2} = Re^{k\arctan\frac{c\cos\varphi+d\sin\varphi}{a\cos\varphi+b\sin\varphi}}$$

$$\underbrace{\arctan\frac{Ry}{Rx} = \arctan\frac{y}{x}}$$

Lemma 4 A power $n \in G$ of an affine transformation maps $\overline{x} \in Re^{k\varphi}$ again to $\overline{x}_1 \in Re^{k\varphi}$.

$$\forall \omega^n \in \mathbb{R}^{2\times2} \text{ and } \overline{x} \in Re^{k\varphi} : \omega^n\overline{x} = \overline{x}_1 \text{ and } \overline{x}_1 \in Re^{k\varphi} \text{ where } n \in G$$

Proof of Lemma 4:
Lemma 3 holds for each point on the log spiral $Re^{k\varphi}$
I.e. also for $\omega\overline{x}$ and for $\omega^2\overline{x}$ and so on ... until $\omega^n\overline{x}$ resp. also for $\omega^{-1}, \omega^{-2}, ..., \omega^{-n}$

Corollary 1 Each power of an affine transformation can be represented as follows.

$$\omega^n = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^n = R_0(\alpha)S(c_x{}^n, c_y{}^n)R_0(-\alpha)R_0(n(-\beta))$$

$$= \begin{pmatrix} c_x^n\cos^2\alpha+c_y^n\sin^2\alpha & (c_x^n-c_y^n)\cos\alpha\sin\alpha \\ (c_x^n-c_y^n)\cos\alpha\sin\alpha & c_y^n\cos^2\alpha+c_x^n\sin^2\alpha \end{pmatrix}R_0(n(-\beta))$$

Proof of Corollary 1:

Because of Lemma 4, Lemma 1 can be also defined for $n \in G$:

$$\omega^n = \psi^n R_0(-\beta)^n \text{ where } \psi = \begin{pmatrix} r & s \\ s & t \end{pmatrix}$$

$$\Rightarrow \omega^n = \left(\psi R_0(-\beta)\right)^n = \psi^n R_0(-\beta)^n = \psi^n R_0(n(-\beta)) =$$
$$R_0(\alpha) \, S(c_x, c_y) \, R_0(-\alpha)R_0(\alpha) \, S(c_x, c_y) \, R_0(-\alpha)R_0(\alpha) \, ... \, S(c_x, c_y) \, R_0(-\alpha)R_0(n(-\beta)) =$$
$$R_0(\alpha) \, S(c_x, c_y) \, R_0(0) \, S(c_x, c_y) \, R_0(0) \, ... \, S(c_x, c_y) \, R_0(-\alpha)R_0(n(-\beta)) =$$
$$R_0(\alpha) \, S(c_x, c_y)^n \, R_0(-\alpha)R_0(n(\alpha+\beta)) = R_0(\alpha) \, S(c_x{}^n, c_y{}^n) \, R_0(-\alpha) \, R_0(n(-\beta))$$

Corollary 2 Lemma 4 can also be defined for powers $n \in \mathbb{R}$ instead of $n \in G$.

$$\forall \omega^n \in \mathbb{R}^{2 \times 2} \text{ and } \overline{x} \in \mathrm{Re}^{k\phi} : \omega^n \overline{x} = \overline{x}_1 \text{ and } \overline{x}_1 \in \mathrm{Re}^{k\phi} \text{ where } n \in \mathbb{R}$$

Proof of Corollary 2:

Lemma 4 also holds for $n \in \mathbb{R}$, if ω^n is represented according to Corollary 1, since powers $n \in \mathbb{R}$ not applicable to ω are now defined for c_x and c_y (i.e. n is no more restricted to G)

Summary (Proof of Theorem 1)

Lemma 4 - using Lemma 3 - shows that ω^n ($n \in G$) generates point sequences on log spirals (the so-called orbits). Corollary 2 points out that for $n \in \mathbb{R}$ the orbits of ω^n form log spirals. Corollary 1 defines - by means of Lemma 1 and Lemma 2 - a representation of $\omega^n = R_0(\alpha) \, S(c_x{}^n, c_y{}^n) \, R_0(-\alpha) \, R_0(n(-\beta))$ which can be used to calculate ω^n in $\mathbb{R} \times \mathbb{R}$ for $n \in \mathbb{R}$, what concludes the chain of the proof for Theorem 1.

Definition of the σ-Hull in $\mathbb{R}^2$

Let $D^*(f, \omega^n)$ be a spiral disk $\bigcup_i^\infty p_i$ with fixed point f, bounded by ω^n with $0 \leq n \leq 2m\pi/\beta$,

$m \in \mathbb{N}$ and radius $|\omega^n(s{-}f)|$, where $p_i \in \mathbb{R} \times \mathbb{R}$ and $d(f, p_i) \leq |\omega^n(s{-}f)|$, $s \in \mathbb{R} \times \mathbb{R}$.

$$\text{Then } D'(f, \omega^n) \equiv \begin{cases} D^*(f, \omega^{-n})^C & n < 0 \\ D^*(f, \omega^n) & n \geq 0 \end{cases} \tag{7}$$

denotes a generalized spiral disk in $\mathbb{R}^2$.

The generalized spiral disc segments described by means of log spirals in $\mathbb{R} \times \phi$ as in (2) can now be defined using a continuos function of powers of affine transformations in $\mathbb{R}^2$:

$$D(f, \omega^n) = D'(f, \omega^n) \cap D'(f, R_0\left(\frac{\pi}{4}\right)\omega^{-n}) \cap D'(f, R_0\left(\frac{\pi}{2}\right)S(-1,1)\omega^n) \tag{8}$$

where $\omega^n = R_0(\alpha)S(c_x{}^n, c_y{}^n)R_0(-\alpha)R_0(n(-\beta))$ and $0 \leq c_x < 1$, $0 \leq c_y < 1$, $-\pi \leq \alpha < \pi$, $-\pi \leq \beta < \pi$
Note: for negative α and β one will get a spiral disk segment reflected by the x-axis

Which - by adapting the definition (3) of the set of fixed points $C_\sigma(S)$ to $\mathbb{R} \times \mathbb{R}$ - leads to the definition of the intersection of generalized spiral disk segments $M_\sigma(X)$ resp. the σ-Hull in $\mathbb{R} \times \mathbb{R}$ instead of $\mathbb{R} \times \phi$ in (4) resp. (5)

$$M_\sigma(X) \equiv \begin{cases} \mathbb{R}\mathrm{x}\mathbb{R} & X = \varnothing \\ \bigcap_{x \in X} D(x, {}^1/|\omega^n(s{-}f)|) & X \neq \varnothing \end{cases} \text{ resp. } \mathcal{H}_\sigma(S) \equiv M_\sigma(C_\sigma(S)) \tag{9}$$

This leads to additional special cases for the σ-Hull defined in $\mathbb{R}^2$:

If $c_x = c_y$ in $R_0(\alpha)S(c_x^n, c_y^n)R_0(-\alpha)R_0(n(-\beta))$

$\Rightarrow D(f, \omega^n)$ is bounded by 2 straight lines forming an angle of $\pi/4$

If $c_x \neq c_y$ and $\alpha = 0$ in $R_0(\alpha) S(c_x^n, c_y^n) R_0(-\alpha) R_0(n(-\beta))$

$\Rightarrow D(f, \omega^n)$ is bounded by 3 exponential parabolas.

Both shapes are necessary to find all orbits (also those laying on non spiral like curves) by which an IFS is bound.

2.3   Calculating Orbits using the σ-Hull in $\mathbb{R}^2$

Calculation of Orbit Generating Transformations

Lemma 5 The curvature proportion ρ of $\omega^n \equiv Re^{k\varphi}$ is defined by a k independent of β

$$\forall \omega^n \equiv Re^{k\varphi} \text{ with } \rho = R/r: \quad k = \frac{\ln(c_x^2 \cos^2 \alpha + c_y^2 \sin^2 \alpha)}{2\arctan\dfrac{(c_x - c_y)\sin \alpha \cos \alpha}{c_x \cos^2 \alpha + c_y \sin^2 \alpha}}$$

Proof of Lemma 5 (cf. Proof of Lemma 3)

Using $\psi \bar{x} = \begin{pmatrix} r & s \\ s & t \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c_x^n \cos^2 \alpha + c_y^n \sin^2 \alpha & (c_x^n - c_y^n)\cos\alpha\sin\alpha \\ (c_x^n - c_y^n)\cos\alpha\sin\alpha & c_y^n \cos^2 \alpha + c_x^n \sin^2 \alpha \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$ of Corollary 1

for the substitution of x and y resp. r and φ in the definition of log spirals:

$$R\sqrt{(a\cos\varphi + b\sin\varphi)^2 + (c\cos\varphi + d\sin\varphi)^2} = Re^{k\arctan\frac{c\cos\varphi + d\sin\varphi}{a\cos\varphi + b\sin\varphi}} \qquad \text{leads to}$$

$$(c_x + c_y\cos^2\alpha)^2 + ((c_x - c_y)\sin\alpha\cos\alpha)^2 = e^{2k\arctan\frac{(c_x - c_y)\sin\alpha\cos\alpha}{c_x\cos^2\alpha + c_y\sin^2\alpha}} \quad \text{for } \varphi = 0$$

$$\Rightarrow 2k\arctan\frac{(c_x - c_y)\sin\alpha\cos}{c_x\cos^2\alpha + c_y\sin^2\alpha} = \ln(c_x^2\cos^2\alpha + c_y^2\sin^2\alpha)$$

Since for fixed α (e.g. tanα $\doteq$ 1) different values of c_x and c_y form log spirals $Re^{k\varphi}$ with arbitrary k (different in ρ) k resp. 1/k can be represented in dependency of c_x and c_y:

$$k = \frac{\ln\left(\frac{c_x^2 + c_y^2}{2}\right)}{2\arctan\frac{c_x - c_y}{c_x + c_y}} \quad \text{resp.} \quad \frac{1}{k} = \frac{\ln\left(\frac{c_x + c_y}{2}\right)}{2\arctan\frac{\sqrt{c_x} - \sqrt{c_y}}{\sqrt{c_x} + \sqrt{c_y}}} \quad \text{(for } 1/k\ c_x, c_y \text{ turn into } \sqrt{c_x} \text{ and } \sqrt{c_y} \text{ !)}$$

Now (A) the nominator resp. (B) the denominator of the above equations for k and 1/k can be equated to $\dfrac{-k\pi}{2(k+1)}$ resp. $\dfrac{-\pi}{2(k+1)}$ to calculate c_x and c_y.

(A) $\Rightarrow c_x^2 + c_y^2 = 2e^{\frac{-k\pi}{2(k+1)}}$ for k and $c_x + c_y = 2e^{\frac{-k^{-1}\pi}{2(k^{-1}+1)}}$ for k^{-1}

(B) $\Rightarrow c_x - c_y = -(c_x + c_y)\tan\dfrac{\pi}{4(k+1)}$ for k and $\sqrt{c_x} - \sqrt{c_y} = -(\sqrt{c_x} + \sqrt{c_y})\tan\dfrac{\pi}{4(k^{-1}+1)}$ for k^{-1}

This leads to:

$$c_x^2 = e^{\frac{-k\pi}{2(k+1)}} \frac{\left(1-\tan\frac{\pi}{4(k+1)}\right)^2}{\left(1+\tan^2\frac{\pi}{4(k+1)}\right)} \quad \text{and} \quad c_y^2 = e^{\frac{-k\pi}{2(k+1)}} \frac{\left(1+\tan\frac{\pi}{4(k+1)}\right)^2}{\left(1+\tan^2\frac{\pi}{4(k+1)}\right)} \quad \text{for k}$$

$$(10)$$

$$\sqrt{c_x} = e^{\frac{-k^{-1}\pi}{2(k^{-1}+1)}} \frac{\left(1-\tan\frac{\pi}{4(k^{-1}+1)}\right)^2}{\left(1+\tan^2\frac{\pi}{4(k^{-1}+1)}\right)} \quad \text{and} \quad \sqrt{c_y} = e^{\frac{-k^{-1}\pi}{2(k^{-1}+1)}} \frac{\left(1+\tan\frac{\pi}{4(k^{-1}+1)}\right)^2}{\left(1+\tan^2\frac{\pi}{4(k^{-1}+1)}\right)} \quad \text{for k}^{-1}$$

Note: Now it obvious why k and 1/k is treated differently, because otherwise at least one of the resulting c_x or c_y can become greater than one in some cases.

The representation of an orbit ω^n by a continuos function of n (Theorem 1) allows to form the σ-Hull for an attractor A paramertized only by c_x and c_y and a fixed α (because in that case an arbitrary β leads to equal curved σ-Hulls according to Lemma 5). That is, to find an orbit generating transformation ω, where a maximal number of discrete boundary points of A lie on the orbit ω^n (for $n \in G$) forming the edge of a σ-Hull (for $n \in R$).

Now generating transformations ω_S of orbits can be calculated using the equivalent representation of a σ-Hull in R^2 if a σ-Hull is found, where at least four boundary points of A p_1, p_2, p_3, p_4 lay on one of the circumscribed spiral segments S with the fixed point f, forming the edge of this (best fitting) hull.

Thus, $\forall p_i$ $i \in \{1,2,3,4\}$: $p_i \in \omega^n \neq p_i \in R_0(\pi/4)\omega^{-n} \neq p_i \in R_0(\pi/2)S(-1,1)\omega^n$ and ω^n forms the edge of the σ-Hull.

The c_x, c_y, α and β - for the log spiral equivalent (Theorem 1) - can be calculated from ω_S by using the representation $\omega = \psi R_0(-\beta)$ and $\psi = R_0(\alpha)S(c_x, c_y)R_0(-\alpha)$ (Lemma 1 and 2).

Optimizing the orbit generating transformations ω_S

$\omega_S = R_0(\alpha)S(c_x^n, c_y^n)R_0(-\alpha)R_0(n(-\beta))$ can now be optimized in the following way :

Since β can also be calculated as $\beta_k = \angle(p_k f R_0(\alpha_k) S(c^{-1}_x, c^{-1}_y) R_0(-\alpha_k)(p_{k+1}))$ and k=1,2,3

$\Rightarrow$ a new $\beta = (\beta_1 + \beta_2 + \beta_3)/3$ can be obtained

Now a new c_x, c_y and α is recalculated using β_k to compute ψ_k:

$\Rightarrow c_x = (c_{1x} + c_{2x} + c_{3x})/3$, $c_y = (c_{1y} + c_{2y} + c_{3y})/3$ and $\alpha = (\alpha_1 + \alpha_2 + \alpha_3)/3$

where $c_{kx} = \frac{r_k + t_k + \sqrt{(r_k - t_k)^2 + 4s_k^2}}{2}$, $c_{ky} = \frac{r_k + t_k - \sqrt{(r_k - t_k)^2 + 4s_k^2}}{2}$ and $\alpha_k = \arctan\frac{c_{kx} - r_k}{s_k}$

for $\psi_k = \omega_S R_0(-\beta_k) = \begin{pmatrix} r_k & s_k \\ s_k & t_k \end{pmatrix} = R_0(\alpha_k) S(c_{kx}, c_{ky}) R_0(-\alpha_k)$ and k as in Lemma 2

Reflection transformations (IFS-Codes) - Either c_x or c_y is negative

If two best fitting σ-Hulls are found which are formed by the log spirals

$\omega_1^{2n} (p_1) = R_0(\alpha) S(c_x^{2n}, c_y^{2n}) R_0(-\alpha) R_0(2n(-\beta))$ or

$\omega_2^{2n+1} (p'_1) = R_0(\alpha) S(c_x^{2n+1}, c_y^{2n+1}) R_0(-\alpha) R_0(-\beta(2n+1))$

for $n \in N^0$ resp. $1/n \in N$ and p_1, $p'_1 \in R^2$, one affine reflective map can be derived :

$\omega_{12} = R_0(\alpha)S(-c_x, c_y)R_0(-\alpha)R_0(-\beta)$ resp. $\omega_{12} = R_0(\alpha)S(c_x, -c_y)R_0(-\alpha)R_0(-\beta)$,

where $c_x < 0$ resp. $c_y < 0$, if $R_0(\gamma)\omega_1^n(p_1) = \omega_2^n(p'_1)$ resp. $R_0(-\gamma)\omega_1^n(p_1) = \omega_2^n(p'_1)$ for $\gamma \geq 0$.

2.4 How to calculate IFS-Codes

Deriving the IFS-Codes from boundary orbits

Form larger and larger curved σ-Hulls, where the c_x and c_y are computed according to Lemma 5 by using smaller and smaller k resp. 1/k (for $Re^{k\phi}$ in the log spiral equivalent of Theorem 1) for a fixed α *(Fig. 2)*.

Each time you get a best fitting hull (at least four boundary points of an attractor A lie on the edge of the σ-Hull) the ω_S - generating that orbit defined by these 4 points - is one IFS-Code of the IFS W_m for a subobject lying outside (i.e. having a part of its boundary in common with the object) as long as $\omega_S \ne \omega_{S'}$ *(Fig. 2)*. I.e. $W_{i+1} = W_i \cup \{\omega_S\}$ for i = $\{1, ..., m\}$, if $\omega_S \ne \omega_{S'}$ where $\omega_{S'} \in W_i$ and $\omega_S \equiv \omega_{S'}$, if $\omega_{S'} = \omega_S \omega_{S''}$ where $\omega_{S''} \in W_i$
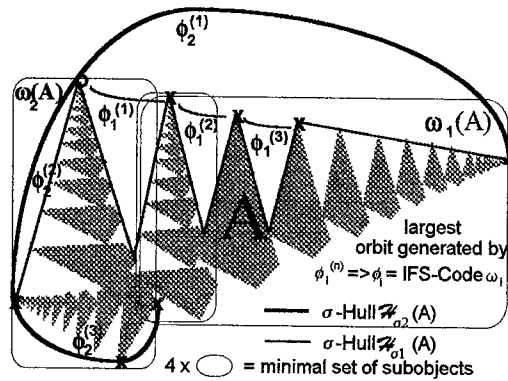
Figure 2. IFS-Code calculation by orbit generating transformations

Besides, this extremely simplifies the classification process (cf. introduction) because larger orbits are found first and the smaller ones have not to be put into the classes any more. Using $\omega^n = R_0(\alpha)S(c_x{}^n, c_y{}^n)R_0(-\alpha)R_0(n(-\beta))$, several special cases can be apriori excluded and thus unnecessary classes are not formed:

- Multiples of other orbits are detected if $S(c'_x, c'_y) = S(c_x{}^n, c_y{}^n)$ n $\in$ G resp. if $\alpha_S = \alpha_{S'}$.
- The orbit generating transformations ω_S of the largest orbits of one class is that ω'_S with the largest β.
- Orbits which can be affinely mapped (by inter orbit transformations) onto eachother will now be detected by those ω_S having a multiple in β - without using an affine invariant representation[7,8].

Thus in the classification process it is not longer necessary to use classes of orbits but classes of the orbit generating and inter orbit transformations represented by their log spiral equivalent (of Theorem 1)

Deriving the IFS-Codes from non boundary orbits

Find all ω_{kj} that for at least one of the spirals S_{1i} (which are formed by each best fitting σ -Hull of A for the IFS-Codes ω_{1i}) circumscribed to $A\backslash\omega^{-1}{}_{kj}(\bigcup^i\omega_{k-1i} A))$ holds $\omega_{kj}(S_{1i})$ =

S_{ki} and S_{ki} a circumscribed spiral to $A\backslash\bigcup^i\omega_{k-1i}(A)$ and ω_{kj} the IFS-Codes for an inner subobject (i.e. having to less boundary points in common with the object).

Repeat this process until $A \setminus \bigcup^{i} \omega_{ki} (A) = \emptyset$, where $\{\omega_{ki}\} = \{\omega_{k-1i}\} \cup \{\omega_{kj}\}$ and $k = \{2,...,n\}$.

Note: Thus the IFS-Codes of attractors with overlapping subobjects can also be calculated.

3   Conclusion

3.1   Summary

In this work is shown how the boundary of a discrete self affine object can be calculated. This is the proposition to form orbits of boundary points of an object. Subsequently the orbits have to be classified to calculate the IFS-Codes (affine maps) of a minimal IFS as generating transformations of the largest of these orbits in each class.

The boundary orbits are found by using a generalized convex hull - the σ-Hull. This hull is formed by the intersection of generalized spiral disk segments (parts of the plane bounded by 3 log spirals equal in curvature proportions and fixed points) instead of half planes used for the convex hull. To reduce the calculation costs, an equivalent representation of spirals (in the polar plane) is defined by powers of affine transformations (in the x/y-plane).

This representation has the additional advantage, that - by the used affine transformations - orbits can easily be calculated for the subsequent classification. Besides, the affine equivalent to log spirals extremely reduces the expense for comparison and differentiation of the orbits within the classification process.

Further more generating transformations for non boundary orbits can be detected after removing the subobjects lying outside which have been already calculated. Thus the IFS-Codes for subobjects having no boundary in common with the object can be computed, even if the subobjects are overlapping.

3.2   Future Work

Implementation

This is a theoretical work, which forms the basis for the IFS-Code calculation and classification resp. overcomes the difficulties of their calculation existing so far. Therefore an implementation will be made together with the elaboration of the classification part which has to be completed to solve the inverse IFS-Problem as a whole.

But the prototype developed for the last work[9] already shows, that in principle this problem is solvable by using a less general hull similar to the σ-Hull and by mapping orbits of boundary points onto eachother.

Besides, the experience made by the implementation of this prototype showed the need to have a more general tool (the σ-Hull) to calculate and classify the orbits as basis for the computation of a minimal IFS.

Analysing IFS-attractors under the aspects of the σ-Hull

So far the affine equivalent of log spirals (the basis of the σ-Hull) has primarily be used to calculate the boundary of an IFS-object. But it can also be an instrument to categorize IFS-objects according to the intrinsic properties, such as structure, location and form of the object parts (subobjects) resp. orbits.

Therefore the correlation between orbits and the parameters of the σ-Hull such as scaling factors (and powers of them) - determining the curvature of the orbits - and the

two angles - one determining the distance of orbit points and the other determining the contraction of an orbit - have to be studied in further detail.

The gained knowledge will also improve the IFS-calculation and contribute to a better adaptation of the σ-Hull.

Can the classification process be more improved by the σ-Hull ?

Some implications of the parameters of the σ-Hull for the classification have been already analysed in this paper, but there should be more implicit properties of the σ-Hull which can be used to classify the orbits.

Besides the choice of the order - not only in dependency of a decreasing function of the curvature proportion - by which the σ-Hulls are formed can influence the discrimination process of the detected orbits to be classified.

Last but not least the affine equivalent to log spirals can be used - within the classification - to decide if one matrix can be divided by another resp. for the comparison of matrices with respect to their fixed points.

Designing a more adequate σ-Hull

Based on the present knowledge about orbits of boundary points of IFS-objects, the basic element of the σ-Hull was designed as the intersection of 3 spiral disks equal in curvature proportion and fixed point. But in praxis it may turn out that more or less log spirals with different curvature and fixed point will do a better job.

This variations also will change the size of the basic element, what will influence the number and accuracy of orbits found on the boundary. That is, roughly speaking how "tight" σ-Hulls can be formed.

So far only two types of basic elements have been investigated. An intersection of two log spiral disks having the same curvature and the same fixed point. But - though simpler to handle - this leads to basic elements infinite in size, which are not sufficient to circumscribe an IFS-object in several cases.

Another basic element was formed by two log spiral disks identically curved in opposite direction and having the same tangent in their two tangent points. This leads to more consistent (with respect to the definition of hulls) mathematical model of the σ-Hull, but this basic element is convex and using it's complement again leads to less "tight" σ-Hulls, which prevent that the fixed points of orbits to lay on the boundary.

Generalization of the σ-Hull

In general each discrete object can be represented by an IFS as long as sufficient many IFS-Codes are used (in the extreme case one IFS-Code per pixel !). Decomposing an object in adequate parts will reduce the number of IFS-Codes[10]. So the question is not to design more and more complex hulls (bounded by highly non linear curves[11]), but to use the σ-Hull to find an minimal number of orbits fitting best into the calculated boundary points. Two possible solutions are outlined below:

One to enlarge the pixel size by an appropriate scale and to fit the orbits into less boundary points. Another to make the pixels smaller - so that each point becomes a boundary point and to find all different orbits within this unconnected point set. The second approach seems to be more complicated (complexity in the point relation), but it solves the encoding problem in one step, because there are no subobjects lying inside, anymore.

References

1. Walter G. Kropatsch & Michael A. Neuhauser, Recovering Iterated Functions, In *Proceedings of Czech Pattern Recognition Workshop'93*, Prenet, Sumava Mountains (November 4-6 1993)
2. Michael F. Barnsley, Fractals Everywhere, *Academic Press* (1988)
3. Michael A. Neuhauser, Diskrete Iterierte Funktionensysteme, Diploma Thesis, *Dept. for Pattern Recognition and Image Processing, Inst. of Automation, Technical University of Vienna* (March 1993)
4. Fisher Y. (Ed.), Fractal Image Compression - Theory and Application, *Springer-Verlag New York* (1995)
5. Edelsbrunner H. & Kirkpatric D.G. & Seidel R., On the Shape of a Set of Points in the Plane, *IEEE Transactions on Information Theory 29(4)* (1983) *pages 551-559*
6. Marcel Worring, Shape Analysis of Digital Curves, In *Ph.D. Thesis*, The *Institute for Informatics of the University of Amsterdam*, Netherlands (June, 1993)
7. Erwin Hocevar & Walter G. Kropatsch, Towards a Better Compression of Self Similar Images (II) - Proofing the Boundary Maxima Matching Algorithm for the IFS Encoding of Affine Self Similar Binary Images, In *Proceedings of ICSP'96*, Beijing, China (October 1996) *pages 1223-1226*
8. John C. Hart, Similarity Hashing: A Model-Based Vision Solution to the Inverse Problem of RIFS, *ASI Workshop on Fractal Image Encoding and Analysis*, Trondheim, Norway (July 1995)
9. Erwin Hocevar & Walter G. Kropatsch, Capturing Self Similarity of Nature into Formulas - A Feature Based Solution for Fractal Compressed Encoding of Monochrome IFS-Objects, *Fractals and Beyond, Editor Miroslav M. Novak, World Scientific Publishing* (1998)
10. Michael F. Barnsley & John H. Elton & D.P. Hardin, Recurrent Iterated Function Systems, In the Journal *Constructive Approximation Vol. 5* (1989) *pages 3-31*
11. Vrscay E.R., Iterated Function Systems: Theory, Applications and the Inverse Problem, In *Proceedings of an ASI Conference on Fractal Geometry and Analysis 1989*, Montreal, Canada, *Kluwer* (1991)

ON THE MULTIFRACTAL PROPERTIES OF PASSIVELY CONVECTED SCALAR FIELDS

J. KALDA

Institute of Cyberentics, TTU, Akadeemia tee 21, 12618 Tallinn, Estonia
E-mail: kalda@ioc.ee

Multifractal spectra are derived for 1- and 2-dimensional cross-sections of passively convected scalar fields; 2- and 3-dimensional single-scale velocity fields in the absence of KAM surfaces are considered. Both the Kraichnan model and real flows with non-zero correlation time are studied. The calculation of $f(\alpha)$-curves is based on the probability density function of the stretching factors of fluid elements. It is shown that strict multifractality holds only for small values of α. New multifractal scalar field — "harmfulness"— is suggested to describe the propagation of environmentally dangerous substances.

Keywords: Turbulent diffusion, passive scalar, intermittence, multifractality.

1 Introduction

The convection of passive scalar by chaotic fluid flow has been studied intensively during the last decades. Special attention has been paid to the analysis of the intermittent structure of scalar fields [1-8]. The scalar density correlation properties have been found to be very far from Gaussian; a clear evidence for this is the multifractal structure of scalar dissipation fields [1-4]

Most of the theoretical studies have been based on the Kraichnan model [5], in which case the velocity field is assumed to be delta correlated in time. Despite of being very idealized, this model is far from trivial and is widely believed to reproduce the most important features of the real passive scalar turbulence. However, this rigorous approach has still been unable to relate the multifractal spectrum $f(\alpha)$ directly to the correlation functions of the velocity field. The main tool for the multifractal analysis has been the generalized Baker map model [which has been also used to calculate the probability density function (PDF) of largest Lyapunov exponents] [2,3]. Recently, new approach has been suggested for the multifractal analysis of the passive scalar dissipation field, which is based on a simple equation describing the stretching of fluid elements [9]. In Sections 2 and 3 we extend the approach to three-dimensional geometry. In Section 4, we define new scalar fields, which are based on the density of the passive scalar, and show that they have also multifractal structure. When the convecting velocity field is oceanic or atmospheric motion, and the passive scalar is an environmentally harmful substance, these fields can be treated as the measures describing the potential damage to the environment.

The parameters of our model equations are defined by the statistical properties of the velocity field. In the case of delta-correlation in time (Kraichnan model), they are directly expressed via the correlation functions. The model is in good agreement with the earlier experimental and theoretical results.

2   Basic equation

The convective diffusion of a passive scalar ϕ is described by the following equation:

$$\frac{\partial \phi}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla}\phi = \kappa \boldsymbol{\nabla}^2 \phi + g. \tag{1}$$

Here g is a source of passive scalar and κ — the molecular diffusivity. We consider a chaotic isotropic single-scale incompressible velocity field $\boldsymbol{v}(\boldsymbol{r},t)$ in the absence of KAM surfaces. Single-scale field is defined as a field for which the Fourier spectrum is constrained into one octave of wavelengths. The characteristic space-scale is chosen to be the unit length. We consider large Peclet number limit, $\kappa \ll \langle |\boldsymbol{v}| \rangle$ and moderate time-scale. More precisely, we assume that the convection has created small-scale structures, but the smallest created scales are still longer than the dissipation scale $\sqrt{\kappa t}$, i.e. $t \lesssim -\ln\kappa/2\gamma$, where γ is the average value of the largest Lyapunov exponent of the velocity field. Then, the stretching of fluid elements and the evolution of passive scalar gradients are related to each other [4,6,10-13]. We study the case when there is no source of dye, and at the initial moment $t = 0$, there was a uniform gradient of dye concentration,

$$g \equiv 0, \qquad \boldsymbol{\nabla}\phi|_{t=0} = \boldsymbol{e}_x, \tag{2}$$

where $\boldsymbol{e}_x$ is the unit vector along x-axis.

With the given assumptions and for 2-dimensional velocity field, the problem of finding the PDF of passive scalar gradients is directly equivalent to the problem of finding the PDF of stretching factors of fluid elements $\rho(r,t) = \exp(h_+ t)\cos\varphi$, where h_+ is the largest Lyapunov exponent and φ — the angle between the respective eigenvector and a fluid element [9]. Indeed, neglecting the seed diffusivity, at a fixed fluid particle, the modulus of the dye gradient $|\boldsymbol{\nabla}\phi|$ scales like the length of an infinitesimal fluid element $\delta r(t)$, initially perpendicular to the gradient $[\delta \boldsymbol{r}(0) \perp \boldsymbol{\nabla}\phi(0)]$: $|\boldsymbol{k}(t)| \propto |\delta \boldsymbol{r}(t)|$. This is due to the incompressibility of the fluid: the fluid parallelogram defined by initially perpendicular vectors $\delta \boldsymbol{r}(t)$ and $\delta \boldsymbol{r}_\perp(t)$ preserves its area $\delta S = |\delta \boldsymbol{r}(t)| \cdot |\delta \boldsymbol{r}_\perp(t)| \sin\alpha$, where α is the angle between the vectors. On the other hand, at the fixed fluid particle, the dye concentration remains unchanged and $|\delta \boldsymbol{r}_\perp(t)| \sin\alpha \propto |\boldsymbol{\nabla}\phi|^{-1}$. Thus, with the proper choice of units and in the absence of seed diffusivity, the stretching factors $\rho = |\delta \boldsymbol{r}(\boldsymbol{r}_0,t)| / |\delta \boldsymbol{r}(\boldsymbol{r}_0,0)|$ and dye gradients are equivalent to each other. The strict equivalence holds for the fluid elements, initially parallel to the isolines of $\phi(\boldsymbol{r},0)$; however, statistically the initial orientation of the fluid elements becomes irrelevant.

Three-dimensional geometry can be treated in a similar way, then the PDF of dye gradients is equivalent to the distribution of stretching factors of fluid surfaces. The argumentation is completely analogous and based on the evolution of fluid parallelepipeds.

The multifractal analysis is based on simple diffusion-convection equation, describing the distribution of fluid elements $l(\rho)$ and fluid surfaces $s(\rho)$ over the stretching factor ρ. We derive it both for the Kraichnan model and for velocity fields of finite correlation time. First we consider the case of real velocity fields of finite correlation time τ. The correlation time is used as the unit time, i.e. $\tau = 1$.

Let us define $l(\rho,t)d\rho$ as the average total length of those pieces of a fluid line, for which $\rho \in [\rho, \rho + d\rho]$ (the length is reduced to the initial length of the fluid line L_0). Then, the PDF of stretching factors and dye gradients is given by $\rho^{-1}l(\rho)$. We consider time increments $\Delta t = 1$, and study the change of the state of fluid elements (fluid surfaces); we neglect the correlation in time for time-scales longer than $\tau = 1$. Let $p(q)dq$ denote the probability of stretching a fluid element by a factor of q. Then we can write $l(\rho, t+1)d\rho = \int p(q)l(\rho/q,t)d(\rho/q)qdq$, or, introducing $\sigma = \ln \rho$ and $\lambda(\sigma,t) = l(\exp \sigma, t)$,

$$\lambda(\sigma, t+1) = \int \lambda(\sigma - \ln q, t)p(q)dq. \tag{3}$$

In the case of 3D velocity fields, the same equation describes the evolution of fluid surfaces, with $\lambda(\sigma,t) = s(\exp \sigma, t)$. However, for a fixed velocity field, the stretching of fluid elements and surfaces are characterized by different functions $p(q)$.

The initial condition (2) implies that the dye gradients are initially delta-distributed. Therefore, in order to keep the equivalence between the stretching coefficients and dye gradients, the initial condition for Eq. (3) should be written as

$$\lambda(\sigma, 0) = \delta(\sigma). \tag{4}$$

This system can be solved by applying the Fourier transform $\lambda(f,t) = \int \lambda(\sigma,t)\exp(-if\sigma)d\sigma$. As a result we obtain

$$\lambda(f,t) = \Pi(f)^t, \Pi(f) = \int p(\exp q)\exp(q - ifq)dq. \tag{5}$$

On the long-time limit, the inverse Fourier transform can be taken via the saddle-point method:

$$\lambda \approx (\partial^2 F/t\partial f^2)^{-1/2}\exp\{F[f_0(h),h]t\}.$$

Here we have denoted $h = \sigma/t$, $F(f,h) = \ln[\int_{-\infty}^{+\infty} p(\exp q)\exp(q + fq)dq] - fh$, and $f_0(h)$ stands for the solution to the equation $\partial F/\partial f = 0$. This result has the same form as that of obtained for the PDF of largest Lyapunov exponents $p(h_m,t)$ via generalized Baker map model [3]. It should be stressed that although the functions $\lambda(h)$ and $p(h_m,t)$ have similar meaning, they are in fact distinct, even asymptotically at $t \to \infty$. This is caused by the contribution of fluid elements, almost perpendicular to the eigenvector of the largest Lyapunov exponent. Note that the function $F[f_0(h),h]$ is defined by the correlation properties of the velocity field. For a typical localized correlation function $p(k)$, function $F(f,h)$ grows linearly at $f \to \pm\infty$. Consequently, $f_0(h)$ is defined only for a finite range of the values of h. Outside of that region, $\lambda \equiv 0$. This is quite a natural result: for real non-Kraichnan flows with finite amplitude of velocity fluctuations, the stretching rate of a fluid element cannot be arbitrarily fast.

For Kraichnan flows and for real flows on the long-time limit, Eq. (5) can be further simplified. Indeed, on long-time limit, the function $\lambda(\sigma)$ has a smooth profile made up of long-wave-length Fourier components. Thus, the function $\Pi(f)$ can be expanded into power series, where only the two first terms are kept, $\Pi(f) \approx$

$-iuf - Df^2$. Here u and D are constants depending on the statistical properties of the velocity field:

$$u = \int p(k) \ln k \, dk, \quad D = \frac{1}{2} \int p(k)(\ln k)^2 dk. \tag{6}$$

Further, the discrete increments become relatively small and can be replaced by time derivative. Then, Eq. (3) can be rewritten as

$$\frac{\partial \lambda}{\partial t} + u\frac{\partial \lambda}{\partial \sigma} = D\frac{\partial^2 \lambda}{\partial \sigma^2}. \tag{7}$$

For the Kraichnan model, this is the exact result, which can be obtained via the Fourier transform of the expression $\lambda(\sigma, t) = \langle \delta[\sigma - \ln |\delta r(t)|] \rangle$. Here $\delta r(t) = \delta r(0) \exp(\int_0^t \boldsymbol{\nabla} v \, dt')$, and the limit $t \to 0$ is to be studied. Then $\Pi(f) \equiv -iuf - Df^2$, with

$$
\begin{aligned}
u\delta(t - t') &= \tfrac{d-1}{4} \langle [v_{r\tau}(t) + v_{\tau r}(t)][v_{r\tau}(t') + v_{\tau r}(t')] \rangle, \\
D\delta(t - t') &= \tfrac{1}{2} \langle v_{rr}(t)v_{rr}(t') \rangle,
\end{aligned}
\tag{8}
$$

where indices r and τ denote the components of the tensor $\boldsymbol{\nabla} v$, and $d = 2, 3$ is the dimensionality of the space. For 3-dimensional stretching of fluid surfaces, $\lambda(\sigma, t) = \langle \delta[\sigma - \ln |\delta S(t)|] \rangle$, where $\delta S(t)^2 = \delta r_1(t)^2 \delta r_2(t)^2 - [\delta r_1(t) \cdot \delta r_2(t)]^2$, and $\delta r_{1,2}(t) = \delta r_{1,2}(0) \exp(\int_0^t \boldsymbol{\nabla} v \, dt')$. Unlike in the case of finite correlation time flows, the values of u and D for 3-dimensional stretching of fluid surfaces are equal to the respective values for fluid elements. This is caused by two circumstances: first, the relevant infinitesimal increments of fluid elements satisfy the incompressibility condition; second, only the symmetric components of the stress tensor are involved. Typically, the stretching of fluid elements is defined by the largest Lyapunov exponent λ_+ (except for small fraction of fluid elements, almost perpendicular to the eigenvector). The stretching of fluid surfaces is governed by the sum $\lambda_+ + \lambda_0$, where λ_0 is the intermediate-valued Lyapunov exponent. Thus, the statistical equality of the two stretching coefficients means that the average value of λ_0 is zero. Finally we note that in the case of the Kraichnan model, the time is measured in arbitrary units, because the correlation time $\tau = 0$.

3 Multifractality of the passive scalar dissipation field

The passive scalar dissipation field, created by turbulent jet has been found to exhibit a multifractal structure [1]. This experimental finding has been addressed in several theoretical studies [2-4]. The analytic results confirm the presence of a multifractal structure. The stretching-coefficient-based approach has been used to calculate the multifractal spectrum $f(\alpha)$ for two-dimensional velocity field [9]. Here we extend this approach to 3-dimensional geometry.

Throughout this section, $\lambda(\sigma, t)$ will be treated as the PDF of dye gradients, i.e. $\sigma = \ln |\boldsymbol{\nabla} \phi|$. We consider the initial conditions (2), which is equivalent to (4). Bearing this in mind, Eq. (7) can be immediately solved,

$$\lambda = (\pi D t)^{-1/2} \exp\left[-\frac{(\sigma - ut)^2}{Dt} \right]. \tag{9}$$

This expression allows us to derive the multifractal spectrum. It is convenient to make use of the pattern formed by fluid curves (or fluid surfaces — for 3-dimensional geometry), which were originally straight lines (plane surfaces), separated by unit length, and perpendicular to the gradient of the dye concentration. Note that the idea of studying a fluid line evolution has been used to calculate the Kolmogorov entropy in 2-dimensional quasi-stationary flow [14].

Experimentally, the multifractality has been observed on 1- and 2-dimensional cross-sections of the 3-dimensional dissipation field [1]. Here we shall study 1 and 2-dimensional cross-sections of 2- and 3-dimensional dissipation fields. We start with 1-dimensional cross-sections, particularly we consider the dependence of the local value of σ on the coordinate ξ along the cross-section. First we note that the characteristic fluctuation amplitude of the dye concentration is 1. Indeed, when the fluid lines (or fluid surfaces) evolve, they will be folded; typically, the density variations of the order of unity are embraced between two approaching each other pieces of the curve. Thus, on the cross-section, the characteristic scale of dye density variations is $\delta \approx 1/|\nabla\psi| \approx \exp(-\sigma)$. The small-scale variations of the function $\sigma(\xi)$ are described by the same scale. However, the function $\sigma(\xi)$ exhibits long-range correlations, as well, because two close each other pieces of a fluid curve are stretched in a similar way. It can be argued that in rescaled coordinates $\zeta = \int^{\xi} \xi' \exp[\sigma(\xi')]d\xi'$, function $\sigma(\zeta)$ is a random Brownian function. Indeed, the distance $\Delta\zeta$ between two fluid particles gives us the estimate, how long (i.e. how many durations of the correlation time) has been that time-period, when these points evolved in an uncorrelated manner. During each correlation time, a fluid element is randomly stressed or stretched leading to the change $\Delta\sigma \approx \pm 1$. Therefore, $\Delta\zeta$ gives us the estimate, how many times are the fluid elements independently stressed or stretched.

Further, the multifractal structure of 1-dimensional cross-sections of 2- and 3-dimensional dissipation fields can be easily analyzed. The overall scalar dissipation in a region $[\xi, \xi+r]$, $r \lesssim 1$ can be estimated as $W_r(\xi) = \int_{\xi}^{\xi+r} \kappa|\nabla\psi|^2 d\xi \approx \kappa k_r^2 \delta_0 = \kappa k_r$, where k_r is the maximal value of $|\nabla\psi|$ over the given region, and $\delta_0 = 1/k_r$. In order to determine the multifractal spectrum $f(\alpha)$, we need to calculate the probability

$$p(r,\alpha) \propto r^{1-f(\alpha)} \tag{10}$$

that the normalized dissipation $w_r(\xi) = W_r(\xi)/W_1(\xi)$ scales as α-th power of r, i.e. $w_r \in [r^{\alpha}, 2r^{\alpha}]$. Substituting the estimate $W_r(\xi) \approx \kappa k_r$, we obtain $W_r(\xi)/W_1(\xi) \approx k_r/k_1$. Then, the probability (10) can be calculated as

$$p(r,\alpha) = \begin{cases} L(\rho_0)r, & L(\rho_0)r \ll 1 \\ \exp[-L(\rho_0)r], & L(\rho_0)r \gg 1 \end{cases}, \rho_0 = k_{01}r^{\alpha}; \tag{11}$$

where $L(\rho_0) = \int_{\rho_0}^{\infty} l(\rho)d\rho$ is the overall length [for 3-dimensional velocity field, this is the overall surface $L(\rho_0) = \int_{\rho_0}^{\infty} s(\rho)d\rho$] per unit area of those parts of the fluid curves (fluid surfaces), which are stretched more than a prefixed factor ρ_0. Indeed, $L(\rho_0)$ is the estimate for the number, how many times a cross-section of unit length is intersected by the fluid curves (fluid surfaces in 3D geometry) of stretching factor

$\rho > \rho_0$. According to (9),

$$L = \frac{1}{2} \exp\left[\left(u + \frac{D}{4}\right)t\right]\left[1 - \operatorname{erf}\frac{\sigma_0 + (u + D/2)t}{\sqrt{Dt}}\right], \tag{12}$$

where $\sigma_0 = \ln \rho_0$. At large values of σ_0, the asymptotics of Eq. (12) is given by $L \approx \exp[-(\sigma_0 + ut)^2/Dt - \sigma_0][1 + 2(\sigma_0 + ut)/Dt]^{-1}$; substituting $\sigma_0 = -\alpha|\ln r| + \ln k_{01}$ we obtain

$$p(r, \alpha) \approx r^{1-\alpha(\sqrt{1+4u/D}-\alpha|\ln r|/Dt)}. \tag{13}$$

Here we have also substituted the value of k_1, which has been calculated by noting that $p(1, \alpha) \approx 1$.

In its strict sense, multifractality assumes that $p(r, \alpha)$ is a power law of r. According to Eq. (13), this is valid only for small values of α, $\alpha \ll Dt\sqrt{1 + 4u/D}/|\ln(r_0)|$, where r_0 is the smallest considered space scale. In that case, expressions (10) and (13) yield

$$f(\alpha) = \alpha\sqrt{1 + 4u/D}. \tag{14}$$

It should be stressed that this expression assumes $Lr \ll 1$, and hence $f(\alpha) \lesssim 1$. On the other hand, slight deviations from multifractality in its strict sense may remain unnoticed when performing numerical schemes of obtaining multifractal spectra. Bearing this in mind, it makes sense to calculate the "effective" multifractal spectrum $f(\alpha)$, which might be obtained in experiments:

$$\langle f(\alpha) \rangle = \ln[p(r_0, \alpha)/r_0]/|\ln r_0|. \tag{15}$$
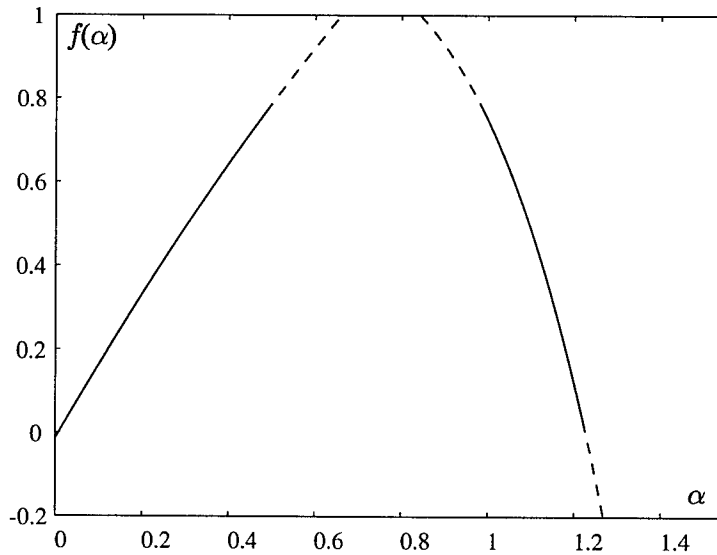

Figure 1. "Effective" multifractal spectrum $\langle f(\alpha) \rangle$, defined by expressions (11), (12) and (15); $u \approx 0.2$, $D \approx 0.35$, $r_0 \approx 0.01$, $t \approx 30$.

Further, we consider 2-dimensional cross-sections of 3-dimensional dissipation field. The cross-sections of the fluid surfaces are curves; in what follows, they will be referred to as "stripes". The variations of the dye gradient are highly anisotropic: the characteristic scale across the "stripes" is previously given by $\delta \approx \exp(-\sigma)$ and is typically defined by the smallest Lyapunov exponent h_-. Meanwhile, the characteristic scale along the "stripes" is much longer and is either defined by the intermediate-valued Lyapunov exponent h_0 (if $h_0 < 0$), or is of the order of unity (if $h_0 > 0$). The easiest way to obtain the $f(\alpha)$-curve is to consider 1-dimensional cross-sections of the 2-dimensional field. The fractal dimension of the intersection of two fractal objects is given by $D_\cap = D_1 + D_2 - D$, where D_1 and D_2 are the dimensionalities of the objects, and D is the topological dimension of the embedding space [15]. Before applying this relationship, it should be pointed out that on 1- and 2-dimensional cross-sections, the same physical points correspond to different values of α. Indeed, on 1-dimensional cross-sections, the normalized dissipation was estimated as k_r/k_1. For 2-dimensional geometry, the striped patterns lead to the estimate $w_r \approx r k_r/k_1 \approx r^\alpha$. By comparing 1- and 2-dimensional expressions for k_r/k_1 (expressed via r and α), one can see that 1-dimensional α corresponds to 2-dimensional $\alpha - 1$ [this correspondence is based on an implicit assumption that the intermediate-valued Lyapunov exponent is not very small, $h_0 t > -\ln(r_0)$]. Therefore, $f(\alpha)$ curves are related by the equality $f_2(\alpha) = 1 + f_1(\alpha - 1)$. Then, expression (14) leads to $f_2(\alpha) = 1 + (\alpha - 1)\sqrt{1 + 4u/D}$.

Finally we note that the curve defined by expressions (11), (12) and (15) is quite similar to the experimental curves, which have been obtained for the passive scalar dissipation in turbulent jet [1]. The leftmost part of the curve is linear, $f \propto \alpha$; this is the only part of the curve corresponding to a strict multifractality. Further, there is a rapid [according to Eq. (11) exponential] fall-off at the large values of α. Exact shape is not reproduced, because expression (11) has been obtained for two asymptotic limits; at the intermediate values of α, the assessment is very rough. Reasonable resemblance between the experimental [1] and theoretical curves is achieved for the following numerical values: $u \approx 0.2$, $D \approx 0.35$, $r_0 \approx 0.01$, $t \approx 30$ (see figure).

4   The measure of "harmfulness".

The convective diffusion is a phenomenon, which will appear in vast number of situations. One of the most important aspects is the mixing of harmful substances in natural environment. These substances may have been delivered as a result of an accident, or as technological wastes; the convecting medium may be oceanic or atmospheric motion. In most cases, the overall amount of the substance is relatively small, so that after complete mixing, there is no danger to the environment. Meanwhile, at the moderate (*under-mixing*) timescale, there are clumps of the substance; biological objects may become damaged, when hit by these clumps. It is not obvious, how to measure the harmfulness of this scalar field. The measures based on the moments of the scalar density are not suited. Indeed, at the under-mixing timescale, the seed diffusivity can be neglected, and one can use the reduced description, where the local state of the medium is given by one bit of information:

the substance is present ('1') or absent ('0'). The local average of the scalar concentration is also useless. This becomes evident, when we consider the evolution of a striped pattern. Due to the incompressibility of the velocity field, the average concentration of the dye will be quasi-homogeneous: in those places, where the stretching factor is high, the stripe becomes very thin; at the same time, it will be approached by another piece of stripe, so that the amount of dye per unit volume will remain unchanged.

Intuitively it is clear that the quantity which does matter, is the size of the blob of dye. Thus, we define new scalar field $\chi(r, t)$ as the radius of the sphere, which is completely immersed into the dye, and the center of which is at the given point. This field — or an n-th power of it $\chi^n(r, t)$ — can be taken as the measure for the harmfulness of the admixture. While $\chi^0(r, t)$ ("zero order harmfulness") is essentially equivalent to the dye concentration itself and is distributed quasi-homogeneously, all the other moments reveal multifractal structure.

In order to study the multifractality of the "harmfulness", we follow the approach of Section 3. Thus, we consider 1-dimensional cross-sections of the scalar field. For the sake of simplicity, let us study only 2-dimensional velocity field (3-dimensional geometry can be handled in the same way, as in the case of dissipation field). Suppose that initially, the stripes of dye of width ϵ were straight lines separated by unit distance. Then, the total n-th order "harmfulness" over a region of size r is estimated as $H_n = \int_\xi^{\xi+r} \chi^n d\xi \approx \epsilon^n \tilde{k}_r^{-n-1}$, where $\tilde{k}_r$ is the smallest value of $|\nabla\psi|$. Note that this estimate is valid for $n > 0$ (actually, for n not too close to 1), because we have taken into account only the contribution of the largest blob of dye in the given region. Strictly speaking, the sum should have been taken over all the blobs. However, just like in the case of dissipation fields, for $n > 0$, the sum is dominated by the largest term. Further, the multifractal spectrum of dissipation fields was derived using the overall length per unit area L of those parts of the fluid curve, which were stretched more than a prefixed factor. Here we have a symmetric situation: the main contribution to the "harmfulness" is made by those parts of the stripes, which have small stretching factors. Therefore, the probability $p(r, \alpha)$ can be assessed by the same formula (11) as in the case of dissipation fields; the only modification is that L should be substituted by $\tilde{L}$ — the length of those parts of the stripes, which are stretched less than $\tilde{k}_1 r^{-\alpha/(n+1)}$ times. Following the procedure of obtaining Eq. (14), we result in

$$f_n(\alpha) = \frac{\alpha}{n+1}\sqrt{1 + 4u/D}. \tag{16}$$

Note that while the multifractality described by Eq. (14) is caused by the "right-hand tail" of the Gaussian distribution function (9) — i.e. by the fall-off at $\sigma \to +\infty$ — the multifractality of the "harmfulness" is due to the left-hand Gaussian tail. Therefore, it is not surprising that Eqns. (14) and (16) are very similar to each other.

5 Conclusions

We have introduced the concept of "harmfulness", which addresses the problem of environmental damage caused by technological wastes. We have shown that in

the case of passive scalar turbulence, both the multifractality of the dissipati field and "harmfulness" are caused by uneven stretching of fluid elements. For multifractality of the dissipation field, this is a known result [2,4]. The advantaε the approach based on Eq. (7) is its simplicity, which made it possible to calcι the $f(\alpha)$-curves. Eq. (7) can be also used to study the other aspects of paε scalar turbulence. Thus, it has been modified to take into account the non-molecular diffusivity [9]. In such a way, it was possible to address the problem stationary $1/k$ power spectrum, and exponential decay of the dye fluctuation the case of initially seeded dye. However, it seems that the approach canno extended to study the PDF of dye gradients in the presence of non-zero diffusiν the dye gradients are affected by the coherence between the existing patterns the patterns formed by diffusing dye.

Acknowledgments

References

1. R. R. Prasad, C. Meneveau, and K. R. Sreenivasan, Phys. Rev. Lett. 6 (1988).
2. E. Ott and T. M. Antonsen, Phys. Rev. Lett. **61**, 2839 (1988).
3. E. Ott and T. M. Antonsen, Phys. Rev. A **39**, 3660 (1989).
4. E. Ott and T. M. Antonsen, Phys. Rev. A **44**, 851 (1991).
5. R. H. Kraichnan, Phys. Fluids **11**, 945 (1968).
6. T. M. Antonsen, Z. F. Fan, and E. Ott, Phys. Rev. Lett. **75**, 1751 (1995)
7. B. S. Williams, D. Marteau, and J. P. Gollub, Phys. Fluids **9**, 2061 (1997
8. R. Pierrehumbert, Chaos Solitons & Fractals **4**, 1091 (1994).
9. J. Kalda Submitted to Phys. Rev. Lett.
10. U. Frisch, A. Mazzino, and M. Vergassola, Phys. Rev. Lett. **80**, 5532 (19ς
11. O. Gat, I. Procaccia, and Reuven Zeitak, Phys. Rev. Lett. **79**, 4577 (1997
12. B. I. Shraiman and E. D. Siggia, Phys. Rev. E **49**, 2912 (1994).
13. M. Chertkov, G. Falkovich, and I. Kolokolov, Phys. Rev. Lett. **80**, 2121 (19
14. A. V. Gruzinov, M. B. Isichenko, and J. Kalda, Sov. Phys. JETP **70**, (1990).
15. B.B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Franc 1983).

NEW STATISTICAL TEXTURAL TRANSFORMS FOR NON-STATIONARY SIGNALS; APPLICATION TO GENERALIZED MULTIFRACTAL ANALYSIS

ANTOINE SAUCIER

*CERCA, 5160 boul. Décarie, bureau 400, Montréal (Québec), Canada H3X 2H9,
e–mail: antoine@cerca.umontreal.ca*

JIRI MULLER

Institutt for Energiteknikk, P.O. Box 40, N-2007 Kjeller, Norway, e-mail: jiri @ife.no

We introduce a method to generate statistical textural transforms that improves the treatment of non-stationarity and leads to a sharper detection of the boundaries between distinct textures (texture segmentation). This method is based on a sliding window processing with fixed size. The basic idea proposed by the authors is to readjust the measuring window around each pixel so as to maximize homogeneity. We use this method with the dimensions $D_n(q)$ that are derived from the Generalized Multifractal Analysis formalism, to show that the $D_n(q)$s can detect and quantify departures from multifractality, while providing the analogue of the classical generalized dimension if the measure is multifractal.

1 Introduction

We propose a method to generate statistical textural transforms that improves the treatment of non-stationarity. Our goal is to improve the quality of the texture segmentations[1] that can be obtained from textural transforms based on a simple sliding window processing[2]. In this context, one usually assumes implicitly that the signal is locally stationary and then proceeds directly with parameter estimation. This assumption is usually not legitimate for every part of the signal. Indeed, there are usually some windows that are not homogeneously textured, which can result in unreliable texture parameters. We propose a more careful treatment of the local homogeneity that leads to significant segmentation improvements. We also apply our method to the generalized multifractal analysis (GMA) representation[3] to show that the generalized dimensions $D_n(q)$ can detect and quantify departures from multifractality, while providing the analogue of the classical generalized dimension $D(q)$ if the measure is multifractal.

2 Non-Stationarity and the legitimacy of statistical texture parameters

For a 1D signal $S(x)$, a resolution preserving[4] *textural transform* associates to each point of the x-axis a number that quantifies an aspect of the local variations of $S(x)$ in the neighborhood of x. The texture parameter associated to a point x_0 is often computed from an interval *centered* on x_0 such as $[x_0 - L_w / 2, x_0 + L_w / 2]$, where L_w is the *window size*. In this case, we talk about a *sliding window processing* of the signal. In the following, the function giving a texture parameter as a function of x will be called a *texture log*. For *statistical* textural transforms, the parameter computed from the signal is statistical in nature.

Texture logs can be used to detect variations in the local statistics of a signal. The signals of interest are therefore typically *non-stationary*. In this paper, we focus on signals for which the textural variations are either slow or abrupt. In other words, we consider signals that are almost *piecewise stationary*, i.e. nearly stationary on consecutive disjoint

intervals. In this context, the presence of discontinuities at the boundaries of adjacent stationary zones raises questions with respect to the sliding window processing. Consider for instance the simple signal shown in Fig. 1. This signal is composed of four segments of equal size (500 points). For each segment, the signal was constructed by adding uniform random deviates (uncorrelated and uniform on [0, 1]) to either a constant or to a slow linear trend. This signal will be regarded as approximately piecewise stationary.

To start with a very simple example of statistical texture parameter, let us suppose that we want to obtain the *mean* texture log of this signal with a sliding window processing, i.e. we compute for each point the arithmetic average of the window centered on this point (Fig. 2). As long as the window lies entirely within one of the four segments, then the data contained in this window is approximately stationary (i.e. statistical *homogeneity* within the window) and it makes sense to compute the mean from the window. However, for windows that *overlap* between two segments, the situation is different. Indeed, such windows then contain two subsets of data that have different and *inconsistent* probability distributions. From a statistical standpoint, it is not legitimate or meaningful to blend two statistically inconsistent samples and then compute a mean. Indeed, the mean obtained *is not representative* of any of the two statistical ensembles. This simple example suggests that *a statistical texture parameter should be computed only for windows that are sufficiently homogeneous statistically*[a], otherwise the resulting parameter is neither representative nor statistically meaningful.

This leads to the problem of measuring the statistical homogeneity of a window. We will use the simple approach that consists in splitting the window in two disjoint sub-samples of equal size, and then to apply a statistical test to compare the distributions obtained from each half. For this comparison, we will use the *Kolmogorov-Smirnov* statistical test that gives the probability *prob1* that the two samples were drawn from the same distribution. To include a sensitivity to 2-point statistics, we will also compare the distributions of the 2-point products $s(i) s(i+n)$ (n = 0, 1, 2, ...) obtained from each half window (here $s(i)$ denotes the value of the signal at point i). For each n, the Kolmogorov-Smirnov test gives a probability *prob2(n)* that the two samples were drawn from the same distribution. We then define a *homogeneity index* by

Homogeneity Index = Min {*prob1*, *prob2 (1)*, *prob2(2)*, ... }

i.e. our homogeneity index is the most pessimistic probability obtained from the statistical comparisons involving one and two point statistics.

In Fig. 2 we plotted the mean texture log together with the homogeneity index for the sample of Fig. 1 (window size = 50 points). The homogeneity index drops to extremely small values at the discontinuities occurring at i = 500 and 1000. Around these locations, the texture log takes the form of a linear transition between two plateaus. Elsewhere, the homogeneity index is large enough to confirm the approximate stationarity of the signal. In our opinion, the windows for which the homogeneity index is very low should not be used to estimate a statistical parameter. For this reason, we propose to revise the idea of a textural transform strictly based on a sliding window processing

---

[a] In this paper, we will use the expressions *statistical homogeneity* and *stationarity* as synonymous.
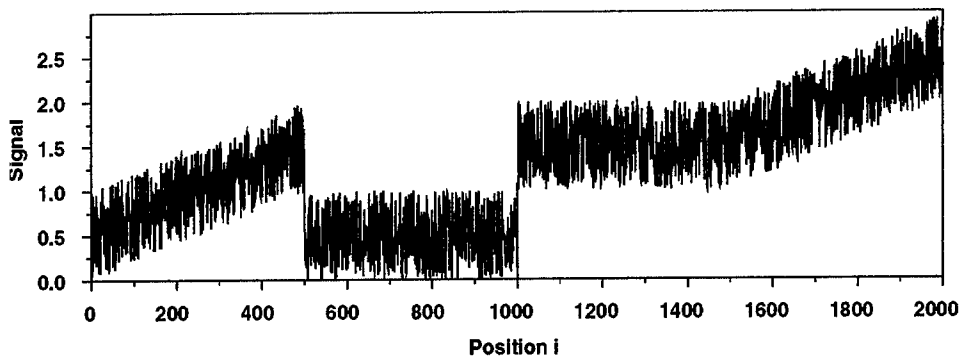
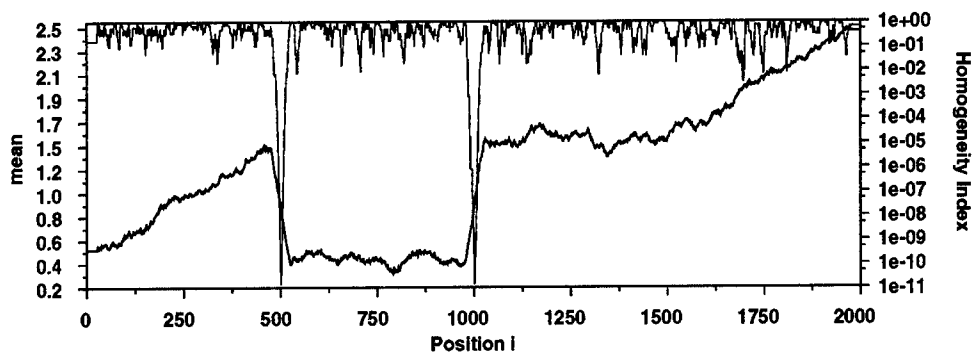Figure 1: An example of a signal that is approximately piecewise stationary.

Figure 2: Mean texture log and homogeneity index (the stalactite looking signal on top).

3 A segmentation-oriented strategy for textural transforms

On one hand, the idea of associating a texture parameter computed from a window to the center of this window is not as natural as it might seem *a priori*. Indeed, *a texture parameter is associated with a window, not with a point*. In reality, any point belonging to this window is equally entitled to receive this texture parameter, especially if every point has contributed equally to its computation. On the other hand, the whole purpose of a texture log is to *detect spatial variations of the local texture*. From this standpoint, it is desirable that the spatial variations of the texture log reflect the spatial variations of the statistics of the underlying signal. Attributing the texture parameter to the window center is a simple way to obtain this dependence.

In this paper, we adopt the standpoint according to which *the textural transform is to be used primarily for segmentation purposes*, i.e. to divide the signal into zones having similar statistics. In other words, the textural transform is regarded as a first step toward segmentation. It follows from this standpoint, for instance, that if there is an abrupt transition between two zones, then the texture log should also exhibit an abrupt transition. It also follows that it is necessary to associate a texture parameter with each point of the signal, i.e. the textural transform should be resolution preserving, which is not possible if only the window centers are considered (i.e. sliding window processing). Finally, a *statistical textural transform* should be based only on windows that are reasonably homogeneous statistically, as far as possible.

To satisfy these guidelines, let us examine again the abrupt transition between two zones, (Figs. 1-2). If the window considered is homogeneous enough, then there is no statistical inconsistency and we choose to adopt the usual sliding window method, i.e. we

attribute to a point x the texture parameter of the window that is centered on x. However, if the window *overlaps* between two different zones A and B, then the window is statistically inconsistent and we want to modify the usual sliding window strategy. If the point x considered belongs to zone A, then it seems natural to attribute to x the texture parameter of a window that both contains x and that lies entirely within the zone A (see point x3 in Fig. 3). Indeed, x belongs to zone A and should therefore be attributed a texture parameter that is representative of this zone. *From this standpoint, the problem is to select for each point x the most appropriate window among all the windows that contain x.*

We propose the following window selection rule for a given point x. Among all the windows that contain x and that have a satisfactory homogeneity (if they exist) we select the window that is most centered on x, i.e. the window for which the distance between its center y and x is minimum. If none of the windows containing x is homogeneous enough, then we simply select the most homogeneous window containing x. Notice that in general there will be two windows at an equal distance $|y - x|$ of x, the left and the right window. To make the choice unambiguous, we select the most homogeneous of these two windows. It is emphasized that if the signal is homogeneous enough, then our definition reduces to the usual sliding window approach. However, if the homogeneity is not acceptable, then our definition forces the windows to "stay away" from discontinuities. We will refer to this approach as the sliding window method with homogeneity correction, or *SWMHC*.

Let us examine a few consequences of our homogeneity correction. If x is close[b] to the beginning of the signal, then the windows containing x cannot be centered on x (see point x1 in Fig. 3). If x is away from the boundaries and in a homogeneous zone, then the window selected will be centered on x (point x2 in Fig. 3). If x is close[c] to a boundary between two zones, then the window selected will tend to be entirely within its own zone (point x3 in Fig. 3).
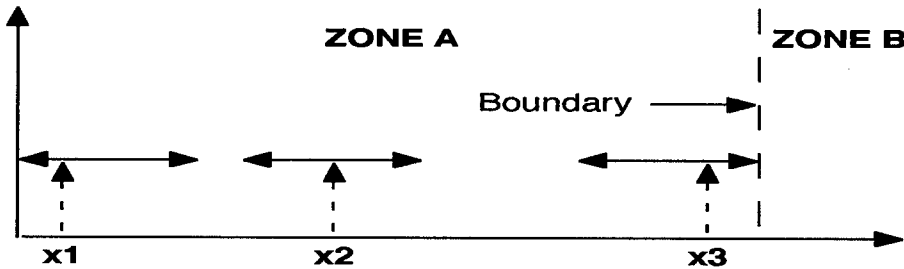
Figure 3: Attribution of windows to each point at the boundary of two zones.

We applied this homogeneity correction algorithm to the signal of Fig. 1, setting the acceptable degree of homogeneity to 0.05 (in statistics, it is usual to reject a hypothesis if the probability associated with a test is smaller than 0.05). It is seen in Fig. 4 that the discontinuities are detected with perfect accuracy, even with a 50-point window size, whereas the rest of the log remains unchanged with respect to Fig. 2. A minimum homogeneity index of 0.05 could be obtained everywhere. Our *SWMHC* has the merit of

---

[b] If the distance between x and the beginning of the signal is smaller than $L_w / 2$, where L_w is the window size.

[c] If the distance between x and the boundary of the signal is smaller than $L_w / 2$, where L_w is the window size.

detecting boundaries sharply even for large windows, which is not possible with the usual sliding window method. Our algorithm reduces uncertainties on texture logs because the most inhomogeneous windows are rejected. In the following, the *SWMHC* will be applied to the generalized dimensions $D_n(q)$.
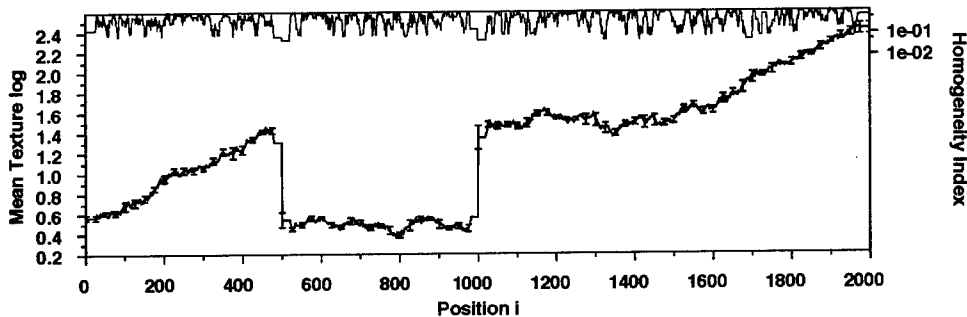

Figure 4: Mean texture log with homogeneity correction. The boundaries at positions 500 and 100 are detected with perfect accuracy, in spite of the fact that a 50-point window size was used.

It is beyond the scope of this paper to give a comprehensive comparison of our method with the numerous other approaches to texture segmentation, ranging from statistical methods, wavelet based methods and neural networks. In the context of wavelet based methods, for instance, clustering procedures[6, 7, 8] or detection of sharp transitions over wavelet energy measurements have been used to produce a final segmentation. These algorithms have good experimental performances but rely on ad hoc parameter settings[9]. The same thing can be said of the segmentation methods that combine statistical parameters and neural networks[10] because the network parameters are adjusted on training sets.

One of the first methods used in texture segmentation, and still a major one, is the spatial gray level co-occurrence matrix[11]. Many authors have approached the problem of texture segmentation with split-and-merge methods combined with co-occurrence matrices[1, 5]. In this context, one creates homogeneous regions by splitting inhomogeneous regions into smaller regions until a given homogeneity criterion is satisfied in each region (the regions can be rectangular if a regular grid is used, or else they can have other shapes depending on the space partitioning method selected). In comparison, our approach consists in constructing a textural transform based on a fixed window size (and shape), but the window containing a given point can be moved *continuously* to satisfy a homogeneity criterion. In addition, our choice of a statistical test to assess homogeneity gives a certain objectivity to the homogeneity criterion.

For a fixed window size, our method cannot always reach an acceptable degree of homogeneity everywhere. However, reducing sufficiently the window size allows reaching an acceptable homogeneity at all points (as in the split-and merge algorithm). An appropriate combination of texture logs obtained for different window sizes can be used to produce a multi-resolution texture log with satisfactory homogeneity everywhere.

4 Brief review of the generalized multifractal analysis representation

4.1 Generalized dimensions defined as projections of the generating function on orthogonal polynomials

We summarize here the generalized multifractal analysis (GMA) representation[3]. The generating function $\chi_q(\delta)$ used in multifractal analysis is a function of two variables q and δ. The parameter q is real and $\delta \geq 0$ is a scale ratio $\delta = \ell / L \leq 1$, where ℓ is a length scale and L is an upper bound on ℓ. Multifractal analysis can be done on the condition that $\chi_q(\delta)$ satisfies the power law property[7] $\chi_q(\delta) \sim \delta^{\tau(q)}$ in an interval $\delta \in [\delta_{min}, \delta_{max}]$ called *scaling range*. With the change of variable $x = -\ln(\delta) \geq 0 \Rightarrow \delta = \exp(-x)$ and the definition $\phi_q(x) \equiv -\ln(\chi_q(\exp(-x)))$, this property can be rewritten in the linear form $\phi_q(x) = \tau_0(q) + \tau_1(q) x$. Many signals cannot be described satisfactorily with this linear model. The GMA representation of $\phi_q(x)$ solves this problem by expanding $\phi_q(x)$ in terms of *orthogonal polynomials* $P_n(x)$ of increasing order n

$$\phi_q(x) = \sum_{n=0}^{N} \tau_n(q) P_n(x) \tag{1}$$

The model (1) is a generalization of the multifractal model. The coefficients $\tau_n(q)$ are generalizations of $\tau(q)^{[7]}$, and $\tau_1(q)$ corresponds to the usual $\tau(q)$ for multifractals. In this paper, we compute $\chi_q(\delta)$ for all the length scales available, i.e. $\ell_n = n, n = 1, 2, ..., N_x$. The coordinates x take discrete values $x_i, i = 1, 2, ..., N_x$ and $\phi_q(x)$ is computed for these $x_i's$. The $P_n(x)$s are orthogonal with respect to a scalar product that we define for any two functions f and g by

$$\langle f, g \rangle \equiv \sum_{i=1}^{N_x} w(x_i) f(x_i) g(x_i) \tag{2}$$

where the $w(x_i) > 0$ are weights. A first possibility is to use *unit weights*, i.e. $w(x_i) = 1$ for each i, which gives a larger weight to large scales. Indeed, the density $\rho(x)$ of the $x_i's$ along the x-axis varies approximately according to $\rho(x_n) = 1/(x_n - x_{n+1}) = 1/\ln(\ell_{n+1}/\ell_n) \approx \ell_n$ and therefore it increases as ℓ_n increases. Another possibility is to use *variable weights* that compensate for the non-uniform density of the $x_i's$ by choosing $w(x_n) = k/\rho(x_n)$ where the constant k is chosen so that $\Sigma_i w(x_i) = 1$. This choice results in approximately uniform weighting along the x-axis in the summation (2). It is this second choice that we make in this paper, mostly because it was found to lead to exponents $\tau_n(q)$ that are more sensitive to textural variations.

The $P_n(x)$s are obtained with a Gram-Schmidt orthogonalization starting with $P_0(x) \equiv 1$ and using iteratively the formula $P_n(x) = x^n - \{\beta_n(0) + \beta_n(1) P_1(x) + ... + \beta_n(n-1) P_{n-1}(x)\}$. Hence they satisfy $\langle P_n, P_m \rangle = \delta_{n,m} \langle P_n, P_n \rangle$ and the $\tau_n(q)$s in (1) are defined by

$$\tau_n(q) = \langle P_n, \phi_q \rangle / \langle P_n, P_n \rangle \tag{3}$$

The $\tau_n(q)'s$ can be always written in the form $\tau_n(q) = (q-1) D_n(q)$, where the $D_n(q)'s$ are formal extensions of the generalized dimensions $D(q)^3$. The linear component $D_1(q)$ reduces to the generalized dimension $D(q)$ if the measure is multifractal, i.e. if $N = 1$ in eq. (1).

4.2 Definition of the generating function

We define the *measure* of an interval $B_r(\delta)$ of size δ centered on a point r by $p_r(\delta) = \Sigma_{r_i \in B_r(\delta)} S(r_i)$. To estimate uncertainties on the generating function, we always consider a collection of M equal size samples ($M \geq 2$) drawn from the same statistical ensemble. For a single window in a signal, we split the window into *four disjoint intervals* of the same size that are regarded as four independent samples. Firstly, we define for each of these samples an "individual" generating function by (this form holds only if $p_r(\delta) > 0$, see [3])

$$\chi_q(\delta) = \delta^{(q-1)D} \langle [p_r(\delta)]^q \rangle_S / (\langle p_r(\delta) \rangle_S)^q \qquad (4)$$

where D is the dimension of the embedding space, and angle brackets $\langle ... \rangle_S$ denote a *spatial* average. The denominator of (4) guarantees that $\chi_q(\delta)$ is normalized exactly, i.e. that $\chi_1(\delta) = 1$ for all δ. Secondly, we define the "global" generating function $\hat{\chi}_q(\delta)$ obtained from the M samples by

$$\hat{\chi}_q(\delta) = \frac{1}{M} \sum_{i=1}^{M} \chi_q^{(i)}(\delta) \qquad (5)$$

where $\chi_q^{(i)}(\delta)$ denotes the individual generating function of sample i. The uncertainty on $\hat{\chi}_q(\delta)$ is estimated by taking into account the variations of $\chi_q^{(i)}(\delta)$ from sample to sample, and the corresponding uncertainties on $D_n(q)$ can be derived[3] (we assume that the $\chi_q^{(i)}(\delta)$s are uncorrelated). The error formulas used in this paper are ($\sigma(X)$ denotes the standard deviation of a random variable X)

$$\sigma^2(\phi_q(x)) = \frac{1}{M} \left(\frac{\sigma(\chi_q)}{\langle \chi_q \rangle} \right)^2 , \quad \sigma^2(\tau_n(q)) = \langle P_n^2, \sigma^2(\phi_q) \rangle / (\langle P_n, P_n \rangle)^2 ,$$

$$\sigma(D_n(q)) = \sigma(\tau_n(q)) / |q - 1|$$

5 Behavior of the $D_n(q)$s for a transition between two distinct random binomial measures

Our goal here is to show that the generalized dimension $D_1(1)$ reduces to the classical information dimension $D(1)$ for such archetypal multifractals. The homogeneity index is computed as previously except for one difference: The two samples obtained from each half of the window are normalized by their mean before the statistical tests are done. This choice is partly justified by the fact that the two halves of a binomial measure are statistically identical only if each half of the measure is individually normalized.

We construct the random binomial measures as usual. Initially, a unit mass is assigned to the interval [0, 1]. In the first step, the unit interval is split in two halves

[0, 1/2] and [1/2, 1]. The first interval receives a random fraction $W_{1,1}$ of the mass, where $0 \leq W_{1,1} \leq 1$, while the second interval receives a mass $1 - W_{1,1}$. In the next steps this splitting procedure is repeated in a self-similar manner: In the step $n+1$, interval i, $[(i-1)(1/2)^n, i(1/2)^n]$ of size $\delta_n = (1/2)^n$ and mass $p_i(n)$, is split in two halves receiving the masses $p_{i,1}(n+1) = W_{n,i} \, p_i(n)$ and $p_{i,2}(n+1) = (1 - W_{n,i}) \, p_i(n)$, where $W_{n,i}$ is a random variable $(1 \leq i \leq 2^n)$. The multipliers $W_{n,i}$ satisfy $0 \leq W_{n,i} \leq 1$ and are identically distributed and mutually independent random variables. We choose here multipliers that take two values w_1 and $w_2 = 1 - w_1$ with equal probability. The signal examined (Fig. 5) is composed of two adjacent binomial measures with parameters $w_1 = 0.25$ and $w_1 = 0.35$, 1024 points each ($n = 10$, i.e. 10 iterations). Each measure was normalized to get a unit root mean square.

We first consider in Fig. 6 the $D_1(1)$ texture log obtained with a window size of 128 points, a maximum box size of 32 points and no homogeneity correction. The dashed lines represent the theoretical values of $D(1)$ for the two binomial measures (we get $D(1) = 0.811$ and 0.934 with $D(1) = -w_1 \log_2(w_1) - w_2 \log_2(w_2)$). $D_1(1)$ fluctuates around the theoretical value of $D(1)$ on both sides, which shows that the GMA representation is consistent with the usual description of multifractals. It can be noticed that $D_1(1)$ oscillates a little below $D(1)$, though the difference is small, which can be explained by the fact that our generating function $\Phi_q(x)$ is not perfectly linear for the binomial measure. The homogeneity index drops to very small values at the boundary between the two measures, located at position 1024. Within each zone, the homogeneity index is relatively large but sometimes drops to small values. This occasional lack of homogeneity results from the fact that the statistical self-similarity of the binomial measure holds exactly only if the two intervals compared with the statistical test match exactly the construction grid of the measure, which is usually not the case.
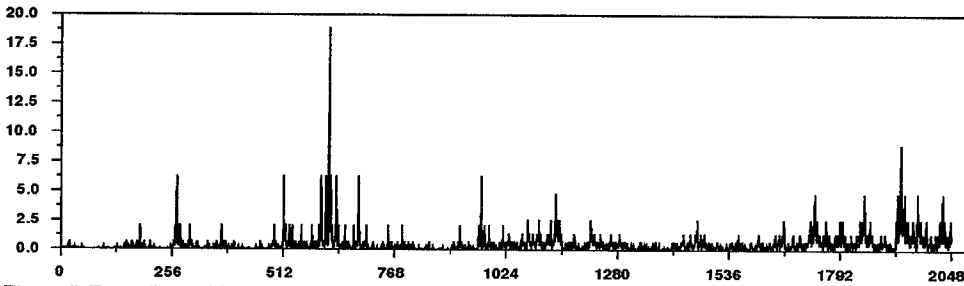

Figure 5: Two adjacent binomial measures.

We recomputed the $D_1(1)$ texture log of Fig. 6 with the homogeneity correction at level 0.05 (Fig. 7). This level was achieved at about 75% of locations, but the level 0.04 is reached almost everywhere. The transition between the two measures is sharp and its location exact, in spite of a window size of 128 points. The amplitude of the fluctuations is reduced by the homogeneity correction. Plateaus are formed, which is a characteristic property of the homogeneity correction. Higher order dimensions, starting with $D_2(1)$, were found to be too small to be distinguished unambiguously (taking into account error bars). This is expected because the function $\Phi_q(x)$ is almost perfectly linear for such binomial measures.
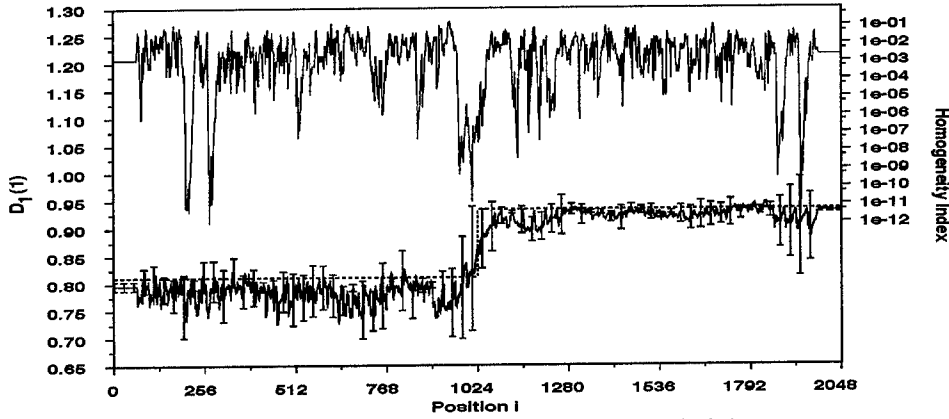
Figure 6: $D_1(1)$ texture log. The stalactite looking curve on top is the homogeneity index.
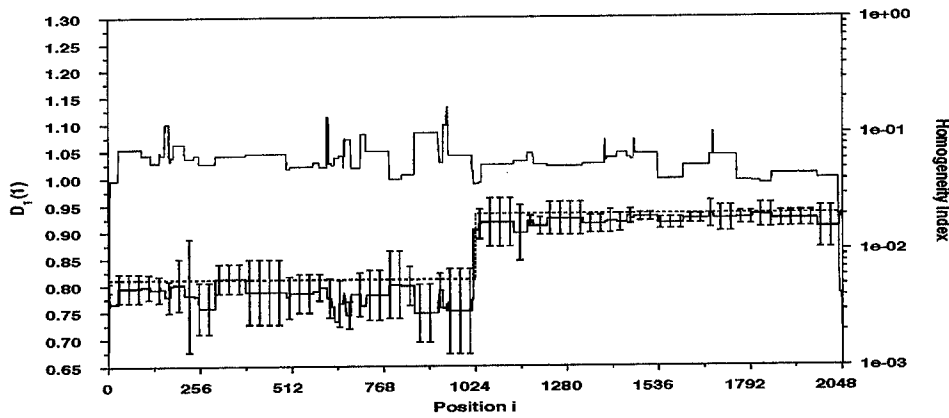

Figure 7: $D_1(1)$ texture log of Fig. 5 with homogeneity correction at level 0.05.

6   Behavior of the $D_n(q)$s for a transition between a binomial measures and a non self-similar measure

In this section, we show that the GMA representation can be used to detect and quantify *departures from self-similarity*, while providing an analogue of the generalized dimension even if the measure is not self-similar. We examine the GMA texture logs for a transition between a random binomial measure with $w_1 = 0.3$ to a *non self-similar* measure. The latter is constructed by making the factor w_1 vary with the step index n (with the multiplicative process described in section 5.1). Our non self-similar measure was obtained with the following values of $w_1(n)$ (n=1,2,...,10): {0.32, 0.27, 0.28, 0.38, 0.25, 0.25, 0.42, 0.40, 0.11, 0.11}. The total measure is obtained by sticking together this measure with the binomial measure (Fig. 8).

The two measures are distinguished clearly by $D_1(1)$ (Fig. 9) and the homogeneity correction again results in a sharp detection of the boundary (Fig. 10). The $D_2(1)$ texture log (Fig. 11) shows a marked transition between the two measures. Indeed, $|D_2(1)|$ is small on the side of the binomial measure (left), and then becomes larger on the side of the non self-similar measure (right). A larger $|D_2(1)|$ indicates a larger deviation of $\Phi_q(x)$ from a straight line, to be expected for a non multifractal.

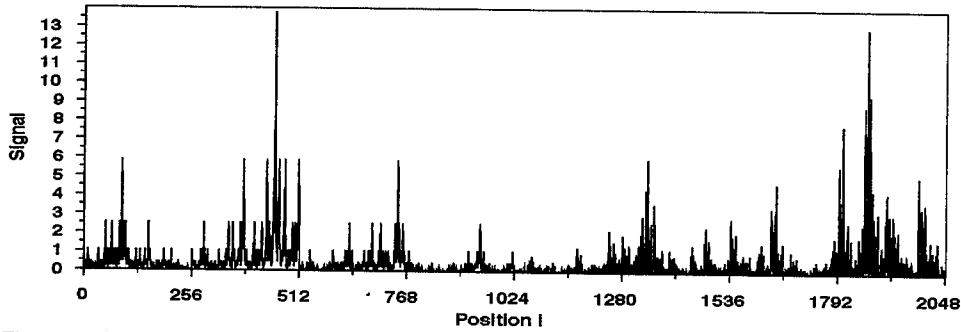**Figure 8**: The first half sample is a random binomial measure with $w_1 = 0.3$. The second half of the sample has been built with a multiplicative process where w_1 varies with the step index.
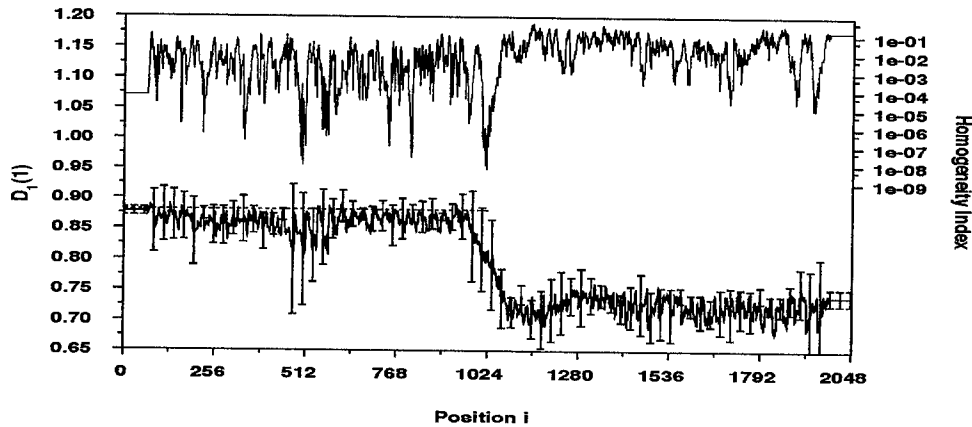

Figure 9: $D_i(1)$ without any homogeneity correction. The dashed line is the theoretical value of $D(1)$.
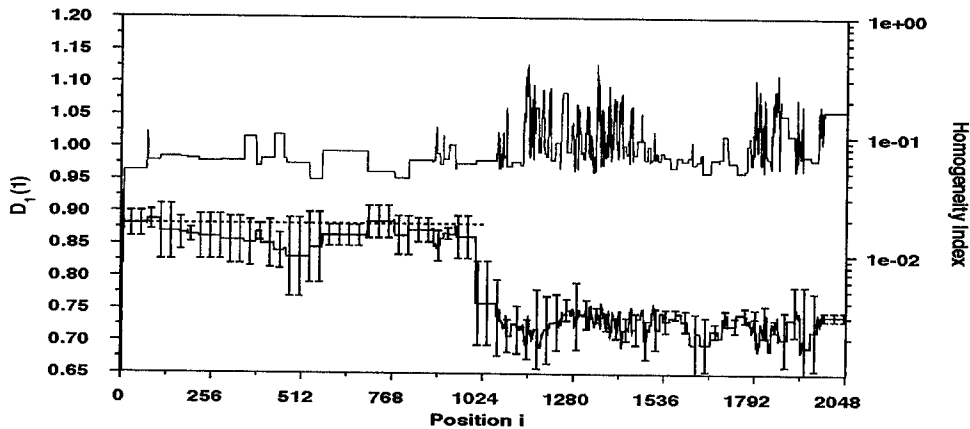

Figure 10: Same than Fig. 10, but with the homogeneity correction at level 0.05.
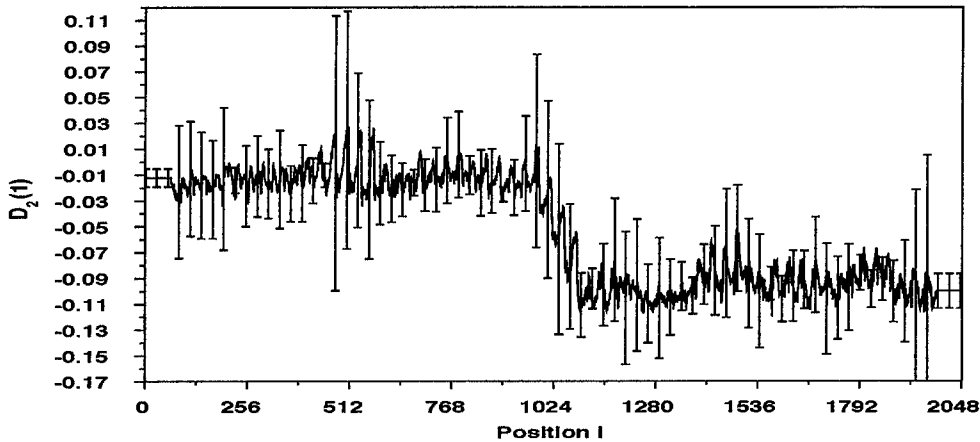
Figure 11: $D_2(1)$ without any homogeneity correction.

7   Conclusions

For signals that are *nearly* piecewise stationary, our method for generating textural transforms results in a sharper detection of the boundary between distinct segments. Indeed, the uncertainty on the location of the transition is much smaller than the window size. These improvements have been obtained in the context of approximately piecewise stationary signals. The special cases examined were the local average of a white noise and the generalized dimensions $D_n(q)$ of adjacent binomial measures. A better assessment of the algorithm proposed should involve segmentation tests in the presence of different types of noise, e.g. colored noise, long-tailed distribution noise, as well as real signals such as 1-d cuts through images.

It is emphasized that our method is quite general and can be used for any statistical parameter. Applied to the GMA representation, we demonstrated that the generalized dimensions $D_n(q)$ can clearly detect departures from multifractality[d] in a signal, while providing the analogue of the classical generalized dimension $D(q)$ for multifractals.

References

1. Haralick R.M., 1986. "Statistical Image Texture Analysis", In: Handbook of Pattern Recognition and Image Processing, Chapter 11, T. Young and K. S. Fu, (eds.), Orlando, Florida: Academic Press, 247--279.

2. A. Saucier, O. K. Huseby, and J. Muller, 1997: Electrical texture characterization of dipmeter microresistivity signals using multifractal analysis. *Journal of geophysical research*, 102:10327-10337.

3. Saucier, A. and J. Muller, 1999. A Generalization of Multifractal Analysis Based on Polynomial Expansions of the Generating Function. In: "Fractals: Theory and Applications in Engineering". Dekking, Lévy Véhel, Lutton, Tricot (eds.), Springer, London. 81-91.

---

[d] Multifractality is used here in the classical sense of a linear behavior of the generating function $\chi_q(\delta)$ plotted as a function of a scale parameter δ d on log-log coordinates.

4. Haralick R.M., 1975. "A Resolution Preserving Textural Transform for Images", Proceedings of the IEEE Computer Society Conference on Computer Graphics, Pattern Recognition, and Data Structure San Diego, California, May 14--16, 51--61.
5. P. C. Chen and T. Pavlidis, 1978. Segmentation by texture using a co-occurrence matrix and a split-and-merge algorithm, *Tech. Rep.* 237, Princeton University, Princeton, New Jersey.
6. A. K. Jain and F. Farrokhnia, 1991. Unsupervised texture segmentation using Gabor filters. *Patt. Recogn.*, 24(12):1167-1186.
7. M. Porat and Y. Zeevi. January 1989. Localized texture processing in vision: analysis and synthesis in Gaborian space. *IEEE Trans. Biomed. Eng.*, 36(1): 115-129.
8. M. Unser, November 1995. Texture classification and segmentation using wavelet frames. IEEE Trans. Image Proc., 4(11):1549-1560.
9. Stéphane Mallat, 1998. A wavelet tour of signal processing. Academic press, London.
10. K. S. Chen, S. K. Yen and D. W. Tsay, 1997. Neural classification of SPOT imagery through integration of intensity and fractal information. *Int. J. Remote Sensing*, 18(4): 763-783.
11. T. R. Reed and J. M. Hans du Buf, May 1993. A Review of Recent Texture Segmentation and Feature Extraction Techniques. *CVGIP: Image Understanding*, 57(3):359-372.

LAPLACIAN GROWTH OF PARALLEL NEEDLES: THEIR MULLINS-SEKERKA INSTABILITY

J.-F. GOUYET AND M.-O. BERNARD

Laboratoire de Physique de la Matière Condensée,
École Polytechnique, F-91128 Palaiseau, France
E-mail: jean-francois.gouyet@polytechnique.fr; marc-olivier.bernard@polytechnique.fr

We study analytically the kinetics of growth of parallel needles, using a conformal transformation to set up the iterative nonlinear equations. This allows to build the discrete Fokker-Planck equation for the probability of finding at time t a given distribution of needle lengths. We consider here two specific cases: we find the exact Fokker-Planck equation for pairs of needles and its solutions, and the linear behavior of a set of n needles with equal initial lengths. The corresponding Fokker-Planck equations show the short-wavelength Mullins-Sekerka instability of these parallel needles, and the possible structure of the screening leading to the scale invariance of the model.

1 Introduction

Laplacian growth is an ubiquitous process of considerable physical interest: start of dendritic growth of crystalline structures during solidification when the diffusion length is large, electrodeposition, viscous fingering, dielectric breakdown or growth of bacterial colonies, belong to Laplacian growth and can be described by Diffusion Limited Aggregation models (DLA)[1]. Related physical phenomena include mechanical cracking, like mud during drying, crack formations in pieces of materials under strain. The Laplacian field takes its origin in the diffusion of material in front of the growing structures, particles which aggregate, nutrient in the growth of bacterial colonies, or in the electric field as in electrodeposition or in dielectric breakdown. In most situations, branching occurs during the growth, leading to DLA structures. In spite of the apparent simplicity of the physical laws governing these systems, and the very simple simulation models containing the essence of the physics, our analytic understanding of DLA remains very unsatisfactory. Essence studies have used extended numerical simulations to extract the scaling laws governing these systems. Therefore a more modest but useful approach consists in a better understanding of DLA growth in the absence of branching[2,3]. Again, in this simpler problem of growing needles, we can consider radial or parallel needles, as well as reflection (*model R*) or absorption (*model A*[4]) of particles (see for instance Krug[5], for a review). In this paper we will concentrate our attention on systems of parallel absorbing needles which, we think, are of more fundamental interest. In two dimensions, conformal mapping allows to obtain analytical results as shown by Derrida and Hakim[6] for radial needles.

An important question in such systems is the competition between the needles. The instability of an initial distribution of equal lengths needles, as well as unstable modes of a growing flat interface in dendritic growth, is known as the "Mullins-Sekerka" instability[7]. In a recent experiment, Losert *et al.*[8] showed the spatial period-doubling instability of dendritic arrays in directional solidification as suggested theoretically by Warren and Langer[9]. It will be the purpose of this

preliminary paper to establish the basic Fokker-Planck equation for the growth of n parallel needles, in the linear regime, which leads to the Mullins-Sekerka behavior. We will indicate in a few words, on the basis of the exact solution for pairs of needles, what we can expect for the general behavior of n parallel needles.

To study analytically the needle models of Laplacian growth in 2-dimensional systems, the most convenient approach is to use conformal transformations. We will only give in Sec. 2, the main arguments to set up these transformations, and the reader is referred to Shraiman and Bensimon[10], Szép and Lugosi[11], Peterson and Furry[12], Kurtze[13], and Derrida and Hakim[6]. For the dynamics, we will follow here Derrida *et al.*[6,14]. In Sec. 3 we write the closed set of equations which describes the time evolution in a compact matrix notation. In Sec. 4, we solve exactly the general two-needles case, and establish the associated Fokker-Planck equation. The case of n growing parallel needles is examined in Sec. 5, and the linearized equations are derived in Sec. 6. These Fokker-Planck equations reveal the short wavelength instability leading to period-doubling.

Two ingredients are needed for the Laplacian growth : on the one hand the Laplacian behavior determines the long range interaction characteristic of DLA growth, but on the other hand the inherent Mullins-Sekerka instability could not operate without the presence of local noise. This noise can be introduced [6,14] in the initial state. We chose here to consider a discrete model with diffusing particles of finite size $\delta\ell$ sticking to the needles at discrete time intervals.

2   The basic conformal transformations

The classical way to parametrize in a convenient manner Laplace's equation with a zero potential boundary condition on a set of n parallel needles is to introduce first a transformation which maps the unit circle in the complex plane z onto an n-branched star in the complex plane ω:

$$\omega = f(z) = Az \prod_{j=0}^{n-1} \left(1 - e^{i\theta_j}/z\right)^{\alpha_j} \quad with \quad \sum_{j=0}^{n-1} \alpha_j = 2. \tag{1}$$

In this transformation, $\pi\alpha_j$ is the angle between two successive needles $\{j-1, j\}$, the sum of the angles being 2π (equ.(1)). The angles θ_j fix the lengths ℓ_i of the needles, and A is a parameter which will be determined below.

A second transformation maps the star in the complex plane ω into a set of parallel needles in a complex plane Ω :

$$\Omega = \log \omega. \tag{2}$$

In the plane Ω, the Laplacian field is $\Phi(\Omega) = Re\left[\log\left(f^{-1}(\exp\Omega)\right)\right]$.

The tip positions in plane z are parametrized by the angles ϕ_i :

$$z_i = \exp(i\phi_i) , \ 0 \leq i \leq n-1 , \tag{3}$$

and the lengths of the needles are the modulus of $\log(f(z))$ at the points $z = z_i$:

$$\ell_i = \log(4A) + \sum_{j=0}^{n-1} \alpha_j \log|\sin((\phi_i - \theta_j)/2)| , \forall \, 0 \leq i \leq n-1 . \tag{4}$$

The n additional constraints must be imposed onto the angles ϕ_i and θ_j which take into account the fact that the needle tips maximize $|f(z)|$ at $z = z_i$:

$$\sum_{j=0}^{n-1} \alpha_j \cot\left((\phi_i - \theta_j)/2\right) = 0 \ , \forall\, 0 \le i \le n-1\ . \tag{5}$$

Now for the kinetics, the growth rate is supposed to be proportional to the potential gradient along the needles. In addition, the growth is supposed to be restricted to the tips. Therefore, following refs.[6,14], where only the tips grow, while the needles remain at zero potential (*model A*), the growth rate of the needles is:

$$\frac{d\ell_i}{dt} \propto \left[\sum_{j=0}^{n-1} \alpha_j \left(1 + \cot^2\left((\phi_i - \theta_j)/2\right)\right)\right]^{-1/2} , \forall\, 0 \le i \le n-1\ . \tag{6}$$

This equation will be used to determine the growth probability of the needles.

Now we have all the ingredients necessary to build the recursion relations to describe the evolution of our system.

3 Matrix form of the basic equations

It appears first very convenient to introduce the matrix $\mathbf{C} = \{c_{ij}\}_{0 \le i,j \le n-1}$ where:

$$c_{ij} = \cot\left((\phi_i - \theta_j)/2\right) \tag{7}$$

and condition (5) may be simply written ($\overrightarrow{\alpha}$ is the vector $\{\alpha_i\}_{0 \le i \le n-1}$):

$$\mathbf{C}.\overrightarrow{\alpha} = \overrightarrow{0}\ . \tag{8}$$

In a similar way, we can define $\mathbf{D} = \{d_{ij}\}_{0 \le i,j \le n-1}$ where $d_{ij} = 1 + c_{ij}^2$. Then condition (6) is :

$$d\ell_i/dt \propto \{\mathbf{D}.\overrightarrow{\alpha}\}_i^{-1/2}. \tag{9}$$

3.1 Infinitesimal growth

We consider now that between t and $t + \delta t$, a particle of size $\delta\ell$ sticks on needle i. This increase of length is supposed much smaller than the distance between needles, $\pi\alpha_j$. The equations can be linearized: for a variation $\delta\ell_i$ of the length ℓ_i corresponds the variations $\delta\phi_i$ of ϕ_i and $\delta\theta_i$ of θ_i. From (4) and (5) we find:

$$\delta\ell_i = \delta A/A - \sum_{j=0}^{n-1} (\alpha_j/2) \cot\left((\phi_i - \theta_j)/2\right) \delta\theta_j\ . \tag{10}$$

As there are $(2n + 1)$ variables in our system of $2n$ equations we have to impose a supplementary constraint. One choice has been to fix one angle (for example $\theta_0 = 0$, ref. [14]). Here we will make the mean of the ℓ_i to remain zero (this choice does not single out any angle). This fixes A via equation (4). Using (10), and $\sum_i \delta\ell_i = 0$,

$$(\delta A/A)\mathbf{1} = 1/(2n)\ \mathbf{O}.\mathbf{C}.\overrightarrow{\alpha\delta\theta}\ , \tag{11}$$

where $\overrightarrow{\alpha\delta\theta} = \{\alpha_i\delta\theta_i\}_{0\leq i\leq n-1}$. Here, $\mathbf{1}$ is the unit matrix and $\mathbf{O} = \{1\}_{ij}$ a matrix filled of 1's. Replacing (11) into (10) gives $\delta\ell_i$ as functions of $\delta\theta_j$,

$$\overrightarrow{\delta L} = -1/2 \ (\mathbf{1} - (1/n)\ \mathbf{O})\,.\mathbf{C}.\overrightarrow{\alpha\delta\theta}\,, \tag{12}$$

where $(\mathbf{1} - (1/n)\ \mathbf{O})$ is a projector on the subspace orthogonal to the completely symmetrical component (for a vector $\overrightarrow{V}$ it is $\widetilde{V}_0 = (1/n)\sum_i v_i$). The choice, $\overrightarrow{\delta L}[j] = {}^t\{-\frac{\delta\ell}{n}, -\frac{\delta\ell}{n}, ..., \{\delta\ell - \frac{\delta\ell}{n}\}_j, ..., -\frac{\delta\ell}{n}\}$ corresponding to adding a particle $\delta\ell$ on needle j, ensures the conservation of the mean length (equal to zero). The ^{t} is for transposition (column vector). Eq. (12) allows to express the variations of $\delta\theta_j$ as functions of the $\overrightarrow{\delta L}$. We can also relate $\delta\phi_i$ to $\delta\theta_j$; Eq. (5) gives :

$$0 = \sum_{j=0}^{n-1} \alpha_j \left(\delta\phi_i - \delta\theta_j\right)\left(1 + \cot^2\left((\phi_i - \theta_j)/2\right)\right)\ . \tag{13}$$

This makes it possible now to calculate the variations $\delta\phi_i$:

$$\delta\phi_i = (\mathbf{D}.\overrightarrow{\alpha\delta\theta})_i/(\mathbf{D}.\overrightarrow{\alpha})_i\ . \tag{14}$$

Now, Eq. (9) which provides the increase of needle length per unit time in the diffusion field, enables us to introduce the growth probability p_i for a particle of size $\delta\ell$ to stick on the tip of needle i during a time δt. Let $(\mathbf{D}.\overrightarrow{\alpha}) = \overrightarrow{\Pi}$,

$$p_i = d\ell_i/(\sum_i d\ell_i)\ = (\Pi_i)^{-1/2}\,/\,(\sum_i (\Pi_i)^{-1/2}). \tag{15}$$

The set of equations, (12), (14) and (15) together with (8) define the successive iterations of the growth: if at time t, all the parameters are known, (15) determines the growth probabilities on the tips; knowing the $\delta\ell_i$ we calculate $\overrightarrow{\delta\theta}$ via (12) then $\overrightarrow{\delta\varphi}$ via (14), and then the perturbation $\delta\mathbf{C}$ of $\mathbf{C}$, and $\delta\mathbf{D}$ of $\mathbf{D}$. To calculate $\overrightarrow{\delta\theta}$ as a function of $\overrightarrow{\delta L}$ we have to invert Eq. (12). Unfortunately, relation (8) shows that $\overrightarrow{\alpha}$ is an eigenvector of $\mathbf{C}$ for the eigenvalue 0, and therefore $\mathbf{C}$ cannot be inverted directly.

3.2 Invariants

In a global rotation $\delta\vartheta$ of the angles θ et ϕ the configuration of the needles and thus all the preceding expressions remain invariant. For example, $\mathbf{C}$ is left unchanged and from (12) and (8) $\overrightarrow{\delta L}$ also remains invariant:

$$\overrightarrow{\delta L} = -(1/n)\ (\mathbf{1} - (1/n)\ \mathbf{O})\,.\mathbf{C}.\left(\overrightarrow{\alpha\delta\theta} + \delta\vartheta\overrightarrow{\alpha}\right) = -(1/n)\ (\mathbf{C} - (1/n)\ \mathbf{O}.\mathbf{C})\,.\overrightarrow{\alpha\delta\theta}\,. \tag{16}$$

4 The two-needles problem

We are now ready to solve the two needles case. Here $n = 2$, but the needles are periodically repeated due to the cyclic boundary conditions imposed as shown in Fig.1. With the zero mean length constraint, we write,

$$\widetilde{\ell}_0 = (\ell_0 + \ell_1)/2 \equiv 0;\quad \widetilde{\ell}_1 = (\ell_0 - \ell_1)/2.$$
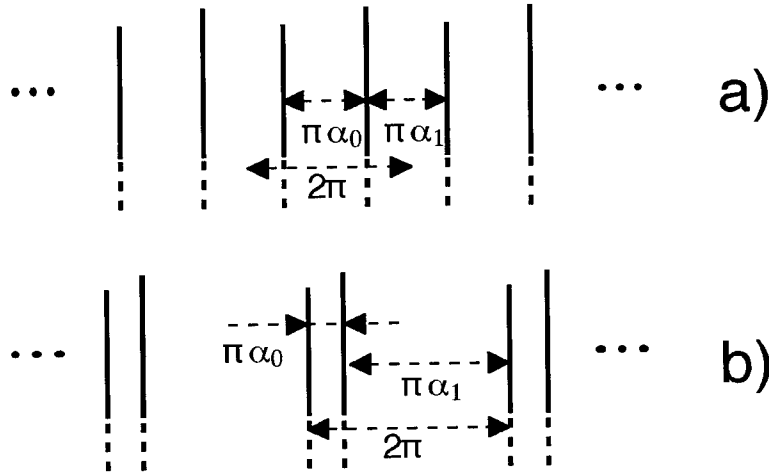
Figure 1. a) Set of equally distant pairs of needles ($\alpha_0 = \alpha_1 = 1$); b) General case.

The problem is completely defined by the knowledge of the probability $P(\widetilde{\ell}_1, t)$ to find at time t a length gap $\widetilde{\ell}_1$. The general form of $\mathbf{C}$, due to condition (8) and of $\mathbf{D}$ with elements $\left(d_{ij} = 1 + c_{ij}^2\right)$ is expressed with two unknown functions $c_0(\widetilde{\ell}_1)$ and $c_1(\widetilde{\ell}_1)$:

$$\mathbf{C} = \left\{ \begin{array}{cc} \alpha_1 c_0 & -\alpha_0 c_0 \\ -\alpha_1 c_1 & \alpha_0 c_1 \end{array} \right\} ; \quad \mathbf{D} = \left\{ \begin{array}{cc} 1 + (\alpha_1 c_0)^2 & 1 + (\alpha_0 c_0)^2 \\ 1 + (\alpha_1 c_1)^2 & 1 + (\alpha_0 c_1)^2 \end{array} \right\} \tag{17}$$

the growth probabilities are then found from (15) :

$$\overrightarrow{p}(\widetilde{\ell}_1) = \left\{ \begin{array}{c} p_0(\widetilde{\ell}_1) \\ p_1(\widetilde{\ell}_1) \end{array} \right\} = \frac{1}{\sqrt{1 + \alpha_0 \alpha_1 c_0{}^2} + \sqrt{1 + \alpha_0 \alpha_1 c_1{}^2}} \left\{ \begin{array}{c} \sqrt{1 + \alpha_0 \alpha_1 c_1{}^2} \\ \sqrt{1 + \alpha_0 \alpha_1 c_0{}^2} \end{array} \right\} \tag{18}$$

4.1 Initial conditions

Fixing arbitrarily $\theta_0(0) = 0$, the initial conditions ($t = 0$), are:

$$\overrightarrow{L}(0) = \left\{ \begin{array}{c} 0 \\ 0 \end{array} \right\} ; \quad \left\{ \begin{array}{c} \theta_0(0) \\ \theta_1(0) \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ \pi \end{array} \right\} ; \quad \left\{ \begin{array}{c} \phi_0(0) \\ \phi_1(0) \end{array} \right\} = \left\{ \begin{array}{c} 2\operatorname{arccot}\sqrt{\alpha_1/\alpha_0} \\ -2\operatorname{arccot}\sqrt{\alpha_1/\alpha_0} \end{array} \right\} \tag{19}$$

and for parameter A,

$$A = (\alpha_0)^{-\alpha_0/2} (\alpha_1)^{-\alpha_1/2} /2. \tag{20}$$

From this we can calculate $\mathbf{C}(0)$ and $\mathbf{D}(0)$.

4.2 Equations satisfied by $c_0(\widetilde{\ell}_1)$ and $c_1(\widetilde{\ell}_1)$

From above, we deduce that the functions $c_0(\widetilde{\ell}_1)$ and $c_1(\widetilde{\ell}_1)$ verify the differential equation:

$$\frac{dc_0(\widetilde{\ell}_1)}{d\widetilde{\ell}_1} + \frac{\left(1 + \alpha_0^2 c_0(\widetilde{\ell}_1)^2\right)\left(1 + \alpha_1^2 c_0(\widetilde{\ell}_1)^2\right)}{\alpha_0\alpha_1\left(1 + \alpha_0\alpha_1 c_0(\widetilde{\ell}_1)^2\right)\left(c_0(\widetilde{\ell}_1) + c_1(\widetilde{\ell}_1)\right)} = 0 \tag{21}$$

and, permuting c_0 and c_1, a similar equation for $c_1(\widetilde{\ell}_1)$. The solutions of these equations can be obtained parametrically:

$$c_0(\widetilde{\ell}_1) = 1/(u\sqrt{\alpha_0\alpha_1}); \quad c_1(\widetilde{\ell}_1) = u/\sqrt{\alpha_0\alpha_1} \tag{22}$$

with the relation,

$$\exp\left[\widetilde{\ell}_1\right] = u\left(\frac{\alpha_0 + \alpha_1 u^2}{\alpha_1 + \alpha_0 u^2}\right)^{(\alpha_0-\alpha_1)/4}. \tag{23}$$

The growth probabilities are then, from (18),

$$p_0(\widetilde{\ell}_1) = u/(1 + u); \quad p_1(\widetilde{\ell}_1) = 1/(1 + u). \tag{24}$$

4.3 Discrete Fokker-Planck equation

The Fokker-Planck equation of a pair of needles is, with $\delta\widetilde{\ell}_1 = \delta\ell/2$,

$$P(\widetilde{\ell}_1, t + \delta t) = p_0\left(\widetilde{\ell}_1 - \delta\widetilde{\ell}_1\right) P(\widetilde{\ell}_1 - \delta\widetilde{\ell}_1, t) + p_1\left(\widetilde{\ell}_1 + \delta\widetilde{\ell}_1\right) P(\widetilde{\ell}_1 + \delta\widetilde{\ell}_1, t). \tag{25}$$

The evolution Eq. (25) can be expanded to second order in $\delta\ell$,

$$\partial P(\widetilde{\ell}_1, t)/\partial t = D.\partial^2_{\widetilde{\ell}_1{}^2} P(\widetilde{\ell}_1, t) - v.\partial_{\widetilde{\ell}_1}\left(P(\widetilde{\ell}_1, t)\,\mathcal{U}(\widetilde{\ell}_1)\right), \tag{26}$$

where the following constants have been introduced,

$$v = \delta\ell/(2\delta t) \text{ and } D = \delta\ell^2/(8\delta t), \tag{27}$$

v is the relative growth velocity of needle 0, D is the "diffusion" coefficient of the sticking between the two needles, and $\widetilde{\ell}_1 = (\ell_0 - \ell_1)/2$. At short time the particles stick at random on both needle tips, up to the moment when one needle gives way to the other. The function $\mathcal{U}$ characterizes the screening effect,

$$\mathcal{U}(\widetilde{\ell}_1) = (u(\widetilde{\ell}_1) - 1)\,/(u(\widetilde{\ell}_1) + 1), \tag{28}$$

where $u(\widetilde{\ell}_1)$ is implicitly defined by (23). The graph of $\mathcal{U}(\widetilde{\ell}_1)$ is shown in Fig.2a.

In the linear region, when $\widetilde{\ell}_1 \to 0$,

$$\mathcal{U}(\widetilde{\ell}_1) \sim \widetilde{\ell}_1/(2\alpha_0\alpha_1) \tag{29}$$

and the diffusion dominates, while when $\widetilde{\ell}_1 \to \pm\infty$, $\mathcal{U}(\widetilde{\ell}_1) \to \pm 1$, and the longest needle grows at velocity v. Eq. (26) represents the diffusion of a "particle" with coordinate $\widetilde{\ell}_1$ in a potential,

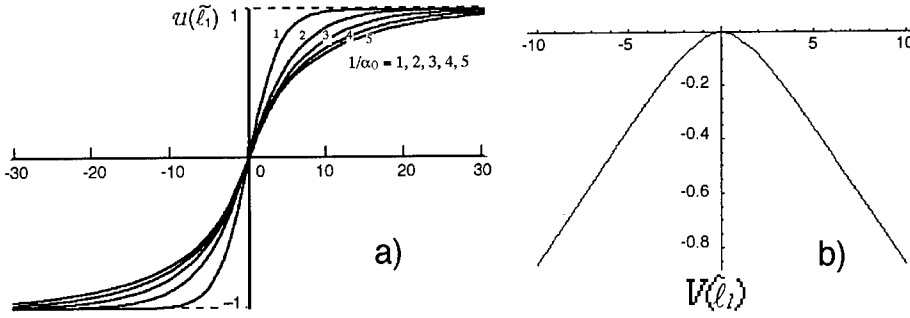$$V(\widetilde{\ell}_1) = -v\int \mathcal{U}(\widetilde{\ell}_1)d\widetilde{\ell}_1. \tag{30}$$

Figure 2. a) Graph of $\mathcal{U}(\widetilde{\ell_1})$ for various values of α_0; b) Effective potential $V(\widetilde{\ell_1})$ when $\alpha_0 = \alpha_1$.
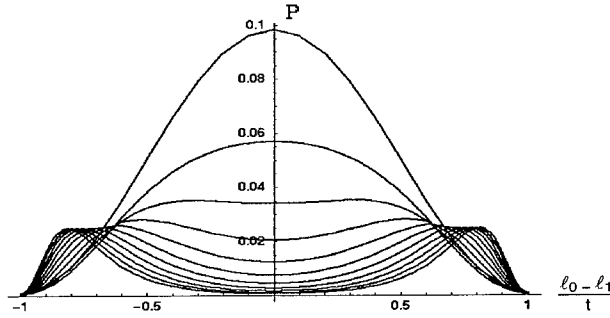

Figure 3. Distribution of probability $P(\widetilde{\ell_1}, t)$ at times $t/\delta t = 20, 30, .., 120$.

If the needles are equidistant, $\alpha_0 = \alpha_1 = 1$, and $u(\widetilde{\ell_1}) = \exp \widetilde{\ell_1}$, we have explicitly, $\mathcal{U}(\widetilde{\ell_1}) = \tanh\left(\widetilde{\ell_1}/2\right)$. The potential (shown in Fig.2b) is then explicitly,

$$V(\widetilde{\ell_1}) = -v \log\left(\cosh(\widetilde{\ell_1}/2)\right). \tag{31}$$

The time evolution of the solution of Eq. (26) for equal length needles $(\alpha_0 = \alpha_1 = 1)$ at $t = 0$, is shown in Fig. 3.

4.4 Fluctuations of the difference between the branch lengths

The average of the difference between the branch lengths is zero by symmetry,

$$\left\langle \widetilde{\ell_1} \right\rangle (t) = \int_{-\infty}^{+\infty} \widetilde{\ell_1} P(\widetilde{\ell_1}, t) d\widetilde{\ell_1} = 0 \tag{32}$$

because $\widetilde{\ell_1} P(\widetilde{\ell_1}, t)$ is odd. But the mean square fluctuation

$$\left\langle \left(\widetilde{\ell_1} - \left\langle \widetilde{\ell_1} \right\rangle\right)^2 \right\rangle (t) = \left\langle \widetilde{\ell_1^2} \right\rangle (t) = \int_{-\infty}^{+\infty} \widetilde{\ell_1^2} P(\widetilde{\ell_1}, t) d\widetilde{\ell_1} \tag{33}$$

is different from zero. Its time evolution is given by the partial differential equation,

$$\frac{\partial \left\langle \widetilde{\ell}_1^2 \right\rangle}{\partial t} = \int_{-\infty}^{+\infty} \widetilde{\ell}_1^2 \frac{\partial P(\widetilde{\ell}_1, t)}{\partial t} d\widetilde{\ell}_1 = \tag{34}$$

$$\int_{-\infty}^{+\infty} \widetilde{\ell}_1^2 \left(D.\partial_{\widetilde{\ell}_1}^2 P(\widetilde{\ell}_1, t) - v.\partial_{\widetilde{\ell}_1} \left(P(\widetilde{\ell}_1, t) \, \mathcal{U}(\widetilde{\ell}_1) \right) \right) d\widetilde{\ell}_1.$$

Integrating by parts leads to,

$$\frac{\partial \left\langle \widetilde{\ell}_1^2 \right\rangle}{\partial t} = 2D + 2v \left\langle \widetilde{\ell}_1 \mathcal{U}(\widetilde{\ell}_1) \right\rangle. \tag{35}$$

When $\widetilde{\ell}_1 \ll 1$, Eq. (35) with (29) reduces to,

$$\frac{\partial \left\langle \widetilde{\ell}_1^2 \right\rangle}{\partial t} \approx 2D + \frac{v}{\alpha_0 \alpha_1} \left\langle \widetilde{\ell}_1^2 \right\rangle, \tag{36}$$

which has the following solution $\left(\left\langle \widetilde{\ell}_1^2 \right\rangle (0) = 0 \right)$ at small enough $\widetilde{\ell}_1$ or t,

$$\left\langle \widetilde{\ell}_1^2 \right\rangle \approx \frac{2\alpha_0 \alpha_1 D}{v} \left(\exp \left(\frac{vt}{\alpha_0 \alpha_1} \right) - 1 \right), \tag{37}$$

The growth behaviour is diffusive $\left(\left\langle \widetilde{\ell}_1^2 \right\rangle \approx 2Dt \right)$, up to a time $t_{dif} \simeq (\alpha_0 \alpha_1 / v)$.

5 Mullins-Sekerka instability of a comb of needles

Let us consider now the case of n equidistant needles for which $\overrightarrow{\alpha} = \frac{2}{n} {}^t\{1, 1, ..., 1\}$. In this case it is convenient to use the Fourier transform of the above equations.

5.1 Reduction of the vector space dimension

We introduce the Fourier transform of an arbitrary n-vector $\overrightarrow{V}$:

$$\overrightarrow{\widetilde{V}} = \mathbf{F}.\overrightarrow{V} \quad \text{with} \quad \mathbf{F} = \{(1/n) \exp(2\pi \mathbf{i} \, jh/n)\}_{0 \le j, h \le n-1} \tag{38}$$

To invert Eq. (12), we have to discard the completely symmetrical component, which makes $\mathbf{C}$ singular. The working space will have now $(n-1)$ components, and we will use an index r to specify its reduced vectors and the rectangular matrix operators relating reduced $(n-1)$-vectors to n-vectors. Thus, we introduce the rectangular Fourier matrix :

$$\mathbf{F}_r = \{F_{kh}\}_{1 \le k \le n-1, \, 0 \le h \le n-1}. \tag{39}$$

The corresponding distribution of modes is,

$$\overrightarrow{\widetilde{L}}_r = \mathbf{F}_r.\overrightarrow{L}, \tag{40}$$

a particular mode k, being given by,

$$\widetilde{\ell}_k = (1/n) \sum_{j=0}^{n-1} \ell_j \exp(2\pi \mathbf{i} \, jk/n). \tag{41}$$

For an increase of length $\delta \ell$ on needle j,

$$\overrightarrow{\delta L_r}[j] = \mathbf{F}_r.\overrightarrow{\delta L}[j] \ = \delta \ell/n \ \{\exp{(2\pi i \ jk/n)}\}_{1 \le k \le n-1} \tag{42}$$

The reduced Fourier matrices lead to the following products: $\mathbf{F}_r.(\mathbf{F}^{-1})_r = \mathbf{1}_r$ and $(\mathbf{F}^{-1})_r.\mathbf{F}_r = \mathbf{1} - (1/n)\mathbf{O}$. In particular, this allows to write (12) as

$$\overrightarrow{\delta L} = -\frac{1}{n} \ (\mathbf{F}^{-1})_r.\mathbf{F}_r. \ \mathbf{C}.\overrightarrow{\delta \theta}. \tag{43}$$

It is now easy to invert Eq. (43). Let $\widetilde{\mathbf{C}}_r = \mathbf{F}_r.\mathbf{C}.(\mathbf{F}^{-1})_r$, then,

$$\overrightarrow{\delta \theta} = -n \ (\mathbf{F}^{-1})_r.(\widetilde{\mathbf{C}}_r)^{-1}\mathbf{F}_r.\overrightarrow{\delta L}. \tag{44}$$

5.2 Evolution of the probabilities

The probability $P(\overrightarrow{L},t)$ to find the distribution of needle lengths $\overrightarrow{L}$ at time t is now replaced by the probability $P(\overrightarrow{L_r},t)$ to find a given Fourier mode, and we have to build the Fokker-Planck equation which relates $P(\overrightarrow{L_r},t)$ and $P(\overrightarrow{L_r}+\overrightarrow{\delta L_r},t+\delta t)$.

With probability $p_i\left(\overrightarrow{L_r}\right)$ we add at time t a particle of size $\delta \ell$ on needle i. The corresponding change of the lengths $\overrightarrow{\delta L}[i]$ induces a change in the angles θ and φ,

$$\overrightarrow{\delta \theta}[i] = -n \ (\mathbf{F}^{-1})_r.(\widetilde{\mathbf{C}}_r)^{-1}\mathbf{F}_r.\overrightarrow{\delta L}[i] \ ; \ \ \overrightarrow{\delta \varphi}[i] = \left\{\left\{\mathbf{D}.\overrightarrow{\delta \theta}[i]\right\}_h \left(\{\mathbf{D}.\overrightarrow{\mathbf{1}}\}_h\right)^{-1}\right\} \tag{45}$$

from which we determine, by differentiation of Eq. (15), the value of $p_i(t+\delta t)$:

$$p_i(t+\delta t) = p_i(t) - \frac{\delta \Pi_i}{2\Pi_i^{3/2}\sum_j \Pi_j^{-1/2}} + \Pi_i^{-1/2}(\ \sum_j \Pi_j^{-1/2})^{-2} \sum_h \frac{\delta \Pi_h}{2\Pi_h^{3/2}}. \tag{46}$$

5.3 Initial conditions

In the present case for which the needles have an equal length at $t=0$, the initial angles ϕ_i et θ_j $(0 \le i,j \le n-1)$ are also regularly spaced :

$$\phi_i(0) = 2\pi i/n \ \ et \ \ \theta_j(0) = \pi(2j-1)/n. \tag{47}$$

5.4 Sticking probabilities in the linearized regime

From expression (45) we can calculate $\overrightarrow{\delta \theta}(0)$ and $\overrightarrow{\delta \varphi}(0)$ corresponding to a variation $\overrightarrow{\delta L}(0)$, for instance by adding a particle on needle h. Thus, if a particle of size $\delta \ell$ is stuck on needle h of a uniform comb, then the sticking probability of the next particle is, using Eq. (46), ($p(t=0) = \frac{1}{n}, D_j = \{\mathbf{D}.\overrightarrow{\mathbf{1}}\}_j$, and $\sum_j D_j^{-1/2} = 1$)

$$p_i = \frac{1}{n} - \frac{\delta D_i(0)}{2n^3} + \frac{1}{2n^4}\sum_{i,j}\delta \mathbf{D}(0)_{i,j}. \tag{48}$$

The sticking probability $p_i[h]$ of a new particle on needle i, while a particle is already stuck on needle h (δ_{ih} is the Kronecker symbol) is then,

$$p_i[h] = \frac{\delta_{ih}}{n} + \delta \ell \left(\frac{(n+1)(n-1)}{6n^2}\delta_{ih} - \frac{\csc^2((i-h)\pi/n)}{2n^2}(1-\delta_{ih})\right). \tag{49}$$
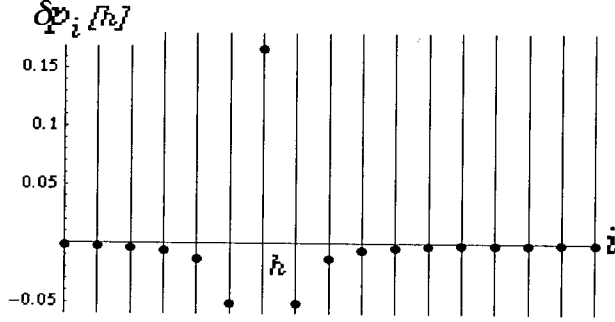
224

Figure 4. Screening of the sticking probability by the longer needle h.

As expected, in this result we see that the other needles are screened by the longer needle h. This is shown in $Fig.4$

6   Fokker-Planck equation of the linearized comb

Let $P_{lin}(\overrightarrow{L}, t)$ be the linearized probability to find a set of needles of size $\overrightarrow{L} = {}^t\{\ell_0, \ell_1, ..., \ell_h, ..., \ell_{n-1}\}$; the linearization supposes that the lengths ℓ_i are not too different in such a way that no needle can be completely screened (see Sec. 4). In the linearized approximation, the sticking probabilities on $\overrightarrow{L}$ are then the superposition of the individual probabilities $p^j[\ell_h]$, where $p^j[\ell_h]$ is deduced from $p_j[h]$ by replacing $\delta\ell$ by ℓ_h (superposition rule). If between t and $t + \delta t$, a particle $\delta\ell$ is added on needle h and if $\overrightarrow{L'_h} = \overrightarrow{L} - \overrightarrow{\delta L}[h]$, then

$$P_{lin}(\overrightarrow{L}, t + \delta t) = \sum_{h=0}^{n-1}\sum_{j=0}^{n-1} p^h[\ell'_j] \ P_{lin}(\overrightarrow{L'}[h], t), \tag{50}$$

with $\overrightarrow{L'}[h] = {}^t\{\ell_0 + \frac{\delta\ell}{n}, \ell_1 + \frac{\delta\ell}{n}, ..., \ell_h - \delta\ell + \frac{\delta\ell}{n}, ..., \ell_{n-1} + \frac{\delta\ell}{n}\}$.
We recover the linearized form of the particular case $n = 2$ of (25):

$$P_{lin}(\widetilde{\ell}_1, t + \delta t) = \left(p^0[\ell'_0] + p^0[\ell'_1]\right) P_{lin}(\widetilde{\ell}_1 - \delta\widetilde{\ell}_1, t) + \left(p^1[\ell'_0] + p^1[\ell'_1]\right) P_{lin}(\widetilde{\ell}_1 + \delta\widetilde{\ell}_1, t) \tag{51}$$

as to first order,

$$p_0\left(\widetilde{\ell}_1 - \delta\widetilde{\ell}_1\right) \simeq (1 + (\ell_0 - \ell_1 - \delta\ell)/4)/2 = p^0[\ell'_0] + p^0[\ell'_1]$$

$$p_1\left(\widetilde{\ell}_1 + \delta\widetilde{\ell}_1\right) \simeq (1 - (\ell_0 - \ell_1 - \delta\ell)/4)/2 = p^1[\ell'_0] + p^1[\ell'_1]. \tag{52}$$

In the general linearized case we obtain from Eqs. (49,50),

$$P_{lin}(\overrightarrow{L}, t + \delta t) = \frac{1}{n}\sum_{h=0}^{n-1} P_{lin}(\overrightarrow{L'}[h], t) + \frac{1}{2n^2}\sum_{h=0}^{n-1}\sum_{j=0, j\neq h}^{n-1} \left(\ell'_h - \ell'_j\right) \csc^2\left((h - j)\pi/n\right) P_{lin}(\overrightarrow{L'}[h], t) \tag{53}$$

with $\overrightarrow{\widetilde{L'}_r}[h] = \overrightarrow{\widetilde{L'}_r} - \overrightarrow{\widetilde{\delta L}_r}[h]$ defined by Eq. (42).

6.1 Fourier transform of the kinetic equation, and the Fokker-Planck equation

Equation (53) can be written after Fourier transformation,

$$
\begin{aligned}
P_{lin}(\overrightarrow{\widetilde{L}}_r, t + \delta t) = \\
\sum_{h=0}^{n-1} \left(\frac{1}{n} - \frac{n^2-1}{6n^2}\delta\ell + \sum_{k=1}^{n-1} e^{-\frac{2\pi i}{n}hk} \frac{k}{n}\left(1 - \frac{k}{n}\right)\widetilde{\ell}_k \right) P_{lin}(\overrightarrow{\widetilde{L}}_r - \overrightarrow{\widetilde{\delta L}_r}[h], t).
\end{aligned}
\tag{54}
$$

In the above equation, we have used the relation,

$$
\sum_{j \neq 0} e^{-\frac{2\pi i}{n}jk} \csc^2(j\pi/n) = (n^2 - 6kn + 6k^2 - 1)/3.
\tag{55}
$$

Second order Taylor expansion of the above Eq. (54) with respect to $\overrightarrow{\widetilde{\delta L}_r}[h]$ makes it clearer. Neglecting the third order terms in $\delta\ell$, we find the remarkable Fokker-Planck equation (note that $\widetilde{\ell}_{-k} \equiv \widetilde{\ell}_{n-k}$, k being defined modulo n),

$$
\begin{aligned}
\frac{\partial P_{lin}(\overrightarrow{\widetilde{L}}_r, t)}{\partial t} = -\frac{\delta\ell}{\delta t} \sum_{k=1}^{n-1} \frac{\partial}{\partial\widetilde{\ell}_k}\left[\frac{k}{n}\left(1 - \frac{k}{n}\right)\widetilde{\ell}_k \, P_{lin}(\overrightarrow{\widetilde{L}}_r, t) \right] \\
+ \frac{\delta\ell^2}{2n^2\delta t} \sum_{k'=1}^{n-1} \frac{\partial^2 P_{lin}(\overrightarrow{\widetilde{L}}_r, t)}{\partial\widetilde{\ell}_{k'}\partial\widetilde{\ell}_{-k'}} + \frac{\delta\ell^2}{2n\delta t} \sum_{k,k'=1}^{n-1} \frac{\partial^2 P_{lin}(\overrightarrow{\widetilde{L}}_r, t)}{\partial\widetilde{\ell}_{k'}\partial\widetilde{\ell}_{k-k'}}.
\end{aligned}
\tag{56}
$$

When $n = 2$, we recover (26) linearized with $\alpha_0 = \alpha_1 = 1$, and $\mathcal{U}(\widetilde{\ell}_1) \sim \widetilde{\ell}_1/2$ from Eq. (29).

6.2 Short discussion of the linearized kinetic equation

Equation (56) contains three terms :

i) the first term is a drift term, associated with the growth velocity v. At the end of the linear regime, we expect an exponential screening behavior similar to $\mathcal{U}(\widetilde{\ell}_1)$ in the 2-needles case. The remarkable point lies in the factor $\frac{k}{n}\left(1 - \frac{k}{n}\right)\widetilde{\ell}_k$. If we start from a situation where all the modes k have the same weight $\widetilde{\ell}_k \equiv \widetilde{\ell}$, then due to the $\frac{k}{n}\left(1 - \frac{k}{n}\right)$ coefficient, the mode $k_{max} = n/2$ dominates progressively since it has the highest growth rate. Due to screening, most of the other modes disappear, and the system becomes equivalent to a comb of $n/2$ needles: a succession of period doubling is the result of the growth.

ii) The second term couples the mode k and $n - k$.

iii) The third term is a mode coupling term, which merges two modes k' and k'' into a mode $k = k' + k''$.

6.3 Correlation between modes and fluctuations of a mode q, in the initial regime

We first calculate the correlation between modes,

$$
\left\langle \widetilde{\ell}_{q_1}\widetilde{\ell}_{q_2} \right\rangle(t) = \int_{-\infty}^{+\infty} \widetilde{\ell}_{q_1}\widetilde{\ell}_{q_2} P(\overrightarrow{\widetilde{L}}_r, t) d\overrightarrow{\widetilde{L}}_r \ \ .
\tag{57}
$$

226

Integration by parts using Eq. (56) gives,

$$\partial_t \left\langle \widetilde{\ell}_{q_1} \widetilde{\ell}_{q_2} \right\rangle_{lin} \Big|_{q_1 \neq q_2} = \frac{\delta\ell}{\delta t} \left(\frac{q_1}{n} \left(1 - \frac{q_1}{n} \right) + \frac{q_2}{n} \left(1 - \frac{q_2}{n} \right) \right) \left\langle \widetilde{\ell}_{q_1} \widetilde{\ell}_{q_2} \right\rangle + \quad (58)$$
$$+ \frac{\delta\ell^2}{2n^2\delta t} \delta_{q_1+q_2,0} + \frac{\delta\ell^2}{2n\delta t} \left(1 - \delta_{q_1+q_2,0} \right).$$

The fluctuation of a mode q, can be also determined from Eq. (56) and leads to,

$$\partial_t \left\langle \widetilde{\ell}_q^{\,2} \right\rangle_{lin} = \frac{2\delta\ell}{\delta t} \frac{q}{n} \left(1 - \frac{q}{n} \right) \left\langle \widetilde{\ell}_q^{\,2} \right\rangle + \quad (59)$$
$$+ \frac{\delta\ell^2}{n^2\delta t} \delta_{q,n/2} + \frac{\delta\ell^2}{n\delta t} \left(1 - \delta_{q,n/2} \right),$$

which has a solution similar to the two-needles case, which underlined the existence of a q-dependent effective velocity,

$$v_{eff}(q) = \frac{2\delta\ell}{\delta t} \frac{q}{n} \left(1 - \frac{q}{n} \right) \quad (60)$$

$$\left\langle \widetilde{\ell}_q^2 \right\rangle_{lin} = \frac{2D_1 \delta_{q,n/2} + 2D_2 \left(1 - \delta_{q,n/2} \right)}{v_{eff}(q)} \left[\exp \left(v_{eff}(q)t \right) - 1 \right] \quad (61)$$

and shows that the mode $q = n/2$ has the fastest growth. We have noted,

$$D_1 = \frac{\delta\ell^2}{2n^2\delta t} \quad \text{and} \quad D_2 = \frac{\delta\ell^2}{2n\delta t}. \quad (62)$$

The correlation between the modes has a similar behaviour,

$$\left\langle \widetilde{\ell}_{q_1} \widetilde{\ell}_{q_2} \right\rangle_{lin} = \frac{D_1 \delta_{q_1+q_2,0} + D_2 \left(1 - \delta_{q_1+q_2,0} \right)}{v_{eff}(q_1, q_2)} \left[\exp \left(v_{eff}(q_1, q_2)t \right) - 1 \right] \quad (63)$$

with an effective velocity,

$$v_{eff}(q_1, q_2) = \frac{\delta\ell}{\delta t} \left(\frac{q_1}{n} \left(1 - \frac{q_1}{n} \right) + \frac{q_2}{n} \left(1 - \frac{q_2}{n} \right) \right) \quad (64)$$

Further studies are in progress to relate the linear regime to the screened regime and to obtain analytically the scaling laws suggested numerically by Krug et al.[4] and Adda Bedia[14].

Acknowledgments

References

1. T.A. Witten, and L.M. Sander, *Phys. Rev. Lett.* **47**, 1400 (1981).
2. G. Rossi, *Phys. Rev.* A **34**, 3543 (1986); *Phys. Rev.* A **35**, 2246 (1987).
3. P. Meakin, *Phys. Rev.* A **33**, 1984 (1986).
4. J. Krug, K. Kassner, P. Meakin, and F. Family, *Europhys. Lett.* **27**, 527 (1993).
5. J. Krug, *Adv. Phys.* **46**, 158 (1997).
6. B. Derrida, and V. Hakim, *Phys. Rev.* A **45**, 8759 (1992).
7. W.W. Mullins, and R.F. Sekerka, *J. Appl. Phys.* **35**, 444 (1964).
8. W. Losert, B.Q. Shi, H.Z. Cummins, and J.A. Warren, *Phys. Rev. Lett.* **77**, 899 (1996).
9. J.A. Warren, and J.S. Langer, *Phys. Rev.* A **42**, 3518 (1990); *Phys. Rev.* E **47**, 2702 (1993).
10. B. Schraiman, and D. Bensimon, *Phys. Rev.* A **30**, 2840 (1984).
11. J. Szép, and U. Lugosi, *J. Phys.* A **19**, L1109 (1986).
12. M.A. Peterson, and J. Ferry, *Phys. Rev.* A **39**, 2740 (1989).
13. D.A. Kurtze, *Phys. Rev.* A **43**, 7066 (1991).
14. M. Adda Bedia, *Ph.D. thesis*, Paris VI University, 1994 (unpublished).

ENTROPY DYNAMICS ASSOCIATED WITH SELF-ORGANIZATION

R. I. ZAINETDINOV

Moscow State University of Railway Communication, 15 Obraztsov Street, Moscow, 101475, Russia
E-mail: zri@hotmail.com

A general model linking dynamics of informational entropy and the self-organization process in an open, steady-state, nonequilibrium system is proposed. Formulas for dynamics of the informational entropy flow and its rate are developed with respect to random process of influences exerted upon a system. It was revealed that the open system responds to a strong change of conditions by steep growth of the informational entropy flow up to a maximum value at the critical point associated with the self-organization process. An example of self-organization during elasto-plastic deformation of metal is considered.

1 Introduction

In this paper, we consider an open, steady-state, nonequilibrium, active system to be a dissipative system. The spectrum of such systems is quite broad. If an exchange of energy and matter with environment is structured, we can designate it as a transfer of information (entropy). In this broad sense, information exchange is not necessarily limited to "intelligent" systems [1]. Presence of an energy flux from an external source to a system and the dissipation of energy on external environment are the preconditions of activity in any system. Because of this, the evolution of open active systems does not necessarily lead towards the equilibrium state. On the contrary, open systems may be involved in processes of self-organization, which result in more complicated and more advanced structures.

As is known [8,9] for open systems the variation of the entropy dS for an interval of time dt can be decomposed into a sum of two components d_eS and d_iS, with quite different physical meanings. d_eS is the entropy flow, which depends on the processes of matter and energy exchange between system and environment. d_iS is the entropy production, caused by irreversible processes inside the system. If conditions $d_eS<0$ and $|d_eS|>d_iS$ are observed, the certain stages of temporal evolution in the open system can occur at a general downturn of entropy $dS<0$.

It is necessary to note that such a situation is possible only far from equilibrium, as in an equilibrium state the member d_iS always prevails. It means that a system is so far from its equilibrium that the linear laws no longer apply; nonlinear terms become important. Self-organization is the "supercritical" phenomenon. Nevertheless far from equilibrium, the system may still evolve to some steady state. In far from equilibrium conditions, various types of self-organization processes may occur.

According to the traditional interpretation of entropy, as a measure of disordering (uncertainty) of a system, it should be considered that, if the disorder decreases at the expense of entropy return in the course of evolution, the system evolves into more complicated and more advanced structures [8].

We present a general model linking dynamics of informational entropy and the self-organization process in an open system. We analyze the local zone's behavior in different "modes of being" [3]. Formulas for dynamics of the entropy flow and its rate are obtained with respect to process of influences exerted upon a system. Finally we describe a few examples in order to elucidate the received results.

2 Definitions and Mathematical Formulation

Let an open system be in a steady (stable), nonequilibrium state. Further, let us assume that in this state the entropy production is compensated for by a negative entropy flow

$$d_i S = -d_e S, \qquad (1)$$

i.e., the system gives back so much entropy as is produced inside the system. Eq. (1) establishes a condition of current balance between the entropy flow through the system and its production inside the system. Current balance is understood as a stationary (not time-dependent) nonequilibrium state of an open system, stable in relation to small deviations. The system in this state is actually in dynamic balance with its environment. The entropy production inside the system in a certain sense characterizes exhaustion of the system's lifespan. Formulas, describing dynamics of the entropy flow and its rate, will be deduced below. According to Eq. (1), these terms can be extended to the internal entropy production inside the system in a steady (stable) nonequilibrium state.

All real processes are irreversible and unbalanced in some degree. Local gradients of temperature, chemical potential, pressure can exist only in a nonequilibrium system. Let X be an important local parameter, determining longevity, lifetime, load-carrying capacity or any other property of vital significance for an open system in a local zone (hereafter we call X the *determining parameter*). Further, let us label a zone of an increased gradient of the determining parameter as the *local dissipative zone* (*LDZ*). Such a local zone limits the lifetime of a system. As a rule, these zones correspond to places of the most probable failures during its lifetime. In the course of intensive operating, the number of LDZ can grow. The growth dynamics of the LDZ number is an important property of the behavior of dissipative system; however, in this paper this question is not considered. Here we investigate the evolution of a system based on processes, occurring only in one LDZ.

There are at least two auto-regulating mechanisms of energy dissipation, which act inside a system. We consider the behavior of a LDZ by using the model of a bistable element. A bistable element (Fig. 1) has two stable states (down and up), in each of which it can exist for a rather long time. Let us denote the mean value of the determining parameter of a system in the down state as X_0 and in up state as X_1. External influences or internal changes in a system can result in transition of the bistable element from one state to another. It was assumed that the transition from down state to up state is caused by such an external influence exerted upon a system, at which determining parameter X (the average over a volume of LDZ) exceeds mean value of the threshold level X_{th}.

In each particular case, the physical content of determining parameter, its threshold level and criterion of transition from one state to another are defined by the type and natures of the system examined and depend on the statement of a problem. In Sec. 5 we shall consider interpretation of this concept for elasto-plastic deformation of metal in the local zone of a load-carrying structure.

For studying the behavior of the LDZ of an open system, the mathematical apparatus of Markovian stochastic processes was used. Solution of the differential Kolmogorov equations for the probabilities $P_0(t)$ and $P_1(t)$ of respectively, the down and up states of the bistable element was obtained [11]. The case of a homogeneous Markovian process were considered, that is the transition rates $\mu(t)$ and $v(t)$ respectively, from the up state to the down state and vice versa do not depend on time: $\mu(t)=\mu$; $v(t)=v$. Hereafter we call $\alpha = v/\mu$ the *regime parameter*. Depending on the regime parameter, α, three typical "modes of being" [3] the LDZ are considered: light at $\alpha<1$, $v<\mu$, $P_0>P_1$, i.e., the LDZ is in
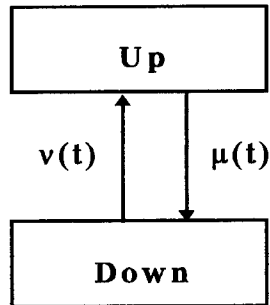
Figure 1. Graph of transition in the bistable element modeling the LDZ behavior.

the down state for a long period of time; symmetric at $\alpha=1$, $v=\mu$, $P_0 = P_1 = 0.5$; heavy at $\alpha>1$, $v>\mu$, $P_1>P_0$, i.e., the LDZ is in the up state for a long period of time.

3 General Results

Let us consider the behavior of the statistical properties of the determining parameter X for the examined LDZ in the course of evolution. Time dependence of mean value X(t) of the determining parameter was determined under the obtained formula

$$\overline{X}(t) = \left[X_0 + \alpha X_1 - \alpha(X_1 - X_0)e^{-\beta t} \right] / (1+\alpha),$$ (2)

where $\beta = v + \mu$. At transient stage, the mean value has the tendency to grow or decrease depending on the initial conditions, as shown in Figure 2 by solid line. Curve 1 occurs when the initial state is the down one and curve 2 – when the initial state is the up. The dashed line corresponds to the steady level of the mean value of determining parameter. One can observe that both lines 1 and 2 approached the steady level at stationary stage.

The second of the principal statistical properties we studied is variance of the determining parameter X. The time course of a variance $D_X(t)$ was described by term

$$D_X(t) = (X_1 - X_0)^2 \left[P_0(t) - P_0^2(t) \right].$$ (3)

The graph of time courses of the variance $D_X(t)$ for heavy and light modes of the system being is shown in Figure 3. Since in the initial moment of time $P_0(t_0)=1$, we have $D_X(t_0)=0$. At the stationary stage (at $t \rightarrow \infty$) we have

$$D_{ST} = \left(X_1 - X_0 \right)^2 \frac{\alpha}{(1 + \alpha)^2}.$$ (4)

For heavy mode of being at $\alpha>1$ in a critical point, corresponding to a moment

$$t_b = -\frac{1}{\beta} \ln\left(\frac{\alpha - 1}{2\alpha} \right)$$ (5)

the function (3) has a maximum. The light mode of the LDZ being when $\alpha<1$ (See line 2 in Figure 3) is characterized by absence of the critical point and by stabilization of the variance on the level D_{ST} in the course of evolution towards a stationary stable state.
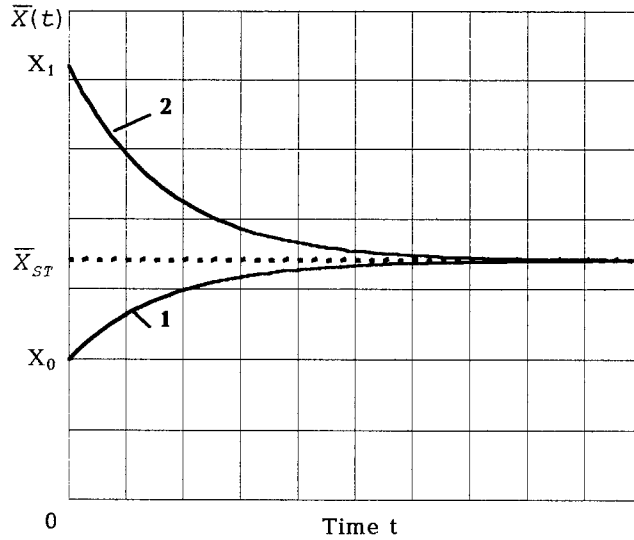
232

Figure 2. The time courses of the mean value of determining parameter (solid lines) from down (1) and up (2) initial states. Steady level is denoted by a dashed line.
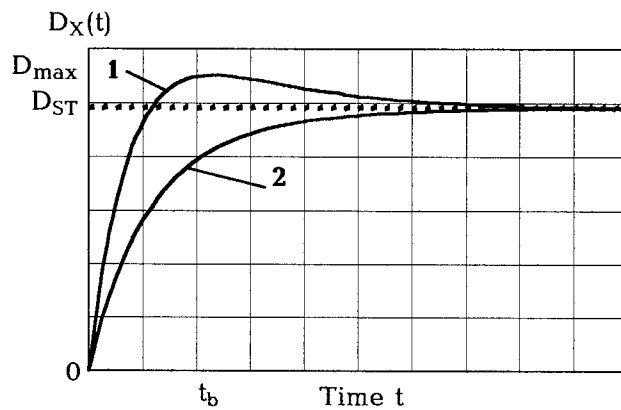

Figure 3. The time courses of the variance $D_X(t)$ for heavy (1) and light (2) modes of being (solid lines). Steady level D_{ST} is denoted by a dashed line. The coordinates of the critical point are labeled as D_{max} and t_b.

Let us assume that the random process $Y(t)$ of dynamic influences with probability distribution function (PDF) $F(Y)$, is exerted upon a system. This process produces a similar random process $X(t)$ of changing the determining parameter X in the LDZ. Further, let us assume the PDF for instantaneous values of the random process is $F(X)$. A dependence of regime parameter α on the PDF of this process is obtained in the form

$$F(X_{th}) = P_0 = 1/(1+\alpha), \quad 1 - F(X_{th}) = P_1 = \alpha/(1+\alpha), \tag{6}$$

where $F(X_{th})$ is the value of PDF $F(X)$ at the threshold level of X_{th}, P_0 and P_1 are final probabilities of the down and up states corresponding to the stationary stage. In Sec. 5 we shall obtain functions connecting regime parameter α and statistical properties (mean value and variance) of normal (Gaussian) random process exerted upon a system.

The main point of interest of the foregoing analysis is the possibility of computing and forecasting the time course of entropy. Note that various interpretations of entropy are internally linked quite closely [7,8]. The physical entropy of a system coincides with the thermodynamic entropy S. The informational entropy H is connected to them by a ratio [7]:

$$S = kH \ln 2, \tag{7}$$

where k is the Bolzmann constant. We consider the informational entropy, being the measure of uncertainty, and equal to the amount of information (according to Shannon) required for removing this uncertainty. For the bistable element modelling the LDZ, the informational entropy H is determined from the equation [7]:

$$H = -\sum_{j=0}^{1} P_j(t) \log_2 P_j(t). \tag{8}$$

The minimum value of entropy $H = 0$ corresponds to the degeneration of a stochastic system into a rigid determinate system. For open self-organizing systems the maximum $H_{max} = 1$ corresponds to a moment of bifurcation, when there is the destruction of pattern (microstructure) exhausting its dissipative abilities, and resulting in an emergence of new pattern at other hierarchical levels. Taking into account decomposition of entropy into a sum of two components and Eq. (1) for current balance, we have considered dynamics of the informational entropy flow in the LDZ of a system and obtained the analytical dependence on of the informational entropy flow on time in the following form:

$$H(t) = -\frac{\alpha}{(1+\alpha)\ln 2}\left\{\frac{1+\alpha e^{-\beta t}}{\alpha}\ln\left(\frac{1+\alpha e^{-\beta t}}{1+\alpha}\right) + \left(1 - e^{-\beta t}\right)\ln\left[\frac{\alpha\left(1 - e^{-\beta t}\right)}{1+\alpha}\right]\right\}. \tag{9}$$

The time courses of the entropy flow H(t) is shown in Figure 4. Analysis of Eq. (9) shows that in heavy mode of the LDZ being (at $\alpha>1$) at the time t_b, determining by Eq. (5), the maximum entropy flow is reached, $H(t_b)=1$. It can be shown that this moment corresponds to the condition of equal probabilities (maximum uncertainty) of keeping the LDZ in the down and up states $P_0(t_b) = P_1(t_b) = 0.5$. Note that this moment t_b corresponds to the critical point of maximum variation D_X^{max} of parameter X. That means, some fluctuations get amplified up to a macroscopic scale. In this way, fluctuations – environmental perturbations or eigenfluctuations – may drive the system into a completely new state and thus become the driving force of system development ("order through fluctuation" [9]). Instabilities at the same time can break symmetry, that is, bifurcations occur: the system may choose among two states, though determined by causality. The selection of the future path of development is unpredictable now [5]. However, at least we can predict the moment of maximum uncertainty and endeavor to take precautions against unfavorable paths of development.

After passing the system through the critical point, which is a stochastic analogue of the bifurcation point [8,12], the entropy flow decreases and, leaving the transient stage, stabilizes on the steady level H_{st}:
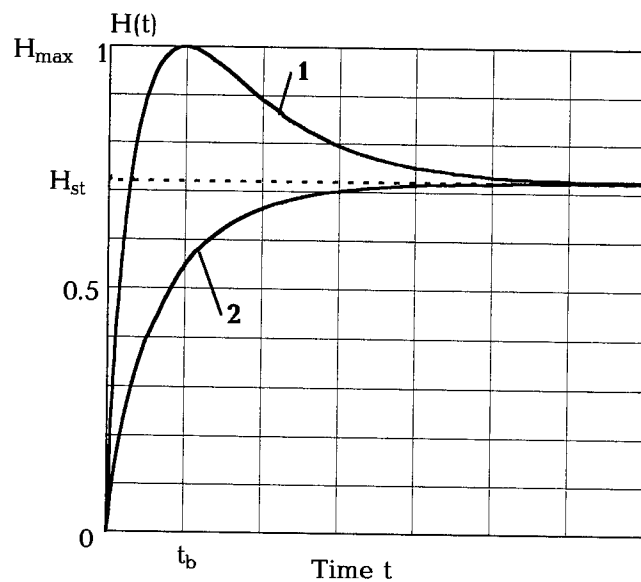
234

Figure 4. The time courses of the informational entropy flow H(t) for heavy (1) and light (2) modes of being (solid lines). Steady level H_{st} is denoted by the dashed line. The coordinates of the critical point are labeled as H_{max} and t_b.
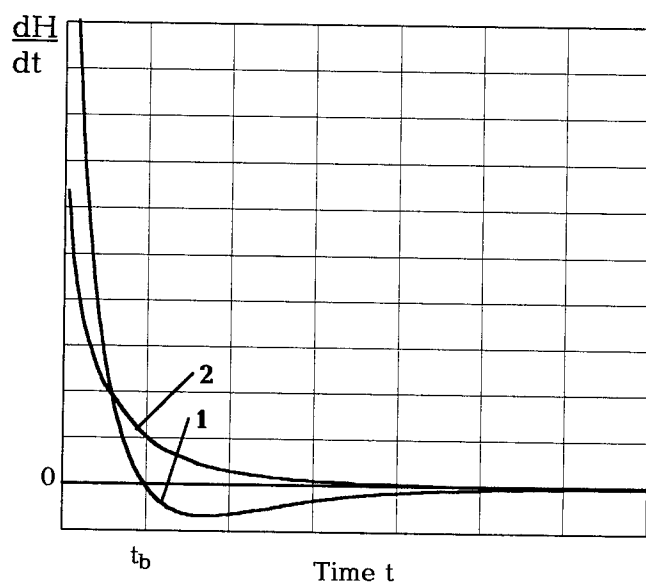

Figure 5. The time courses of the informational entropy flow rate dH/dt for heavy (1) and light (2) modes of being.

$$H_{st} = \ln\left[(1+\alpha)\bigg/ \alpha^{\frac{\alpha}{1+\alpha}}\right] / \ln 2 .$$ (10)

According to the traditional interpretation of entropy, it means that because of entropy outflow in the course of evolution, the disorder decreases. A system is structured in response to the heavy being mode by self-organization of more complicated and more advanced patterns. This is an attribute of the system's adaptation to the random process characterized by the regime parameter $\alpha > 1$.

The light being mode of the LDZ when $\alpha < 1$ (See line 2 in Figure 4) is characterized by absence of the stochastic analog of the bifurcation point and by stabilization of the entropy flow on the level H_{st} during the period of exit from the transient to stationary stage.

Special interest is attracted by dynamics of the entropy flow rate in the course of the system's evolution. The rate is defined as the derivation of the entropy flow function as:

$$\frac{dH}{dt} = -\frac{ve^{-\beta t}}{\ln 2}\ln\left[\frac{\alpha\left(1 - e^{-\beta t}\right)}{1 + \alpha e^{-\beta t}}\right].$$ (11)

Plot of the entropy flow rate with time is shown in Figure 5. Analyzing the function (11) allows us to conclude that the response of a system to the heavy being mode ($\alpha > 1$) by a rapid increase of the entropy flow takes place simultaneously with the reduction of the entropy flow rate to zero at the moment of time t_b. Hereafter the rate of entropy flow becomes negative, passes through a minimum, and aspires to zero, when the transient process approaches the steady (stable) stage.

4   Influence of the Variation of Conditions

Consider now the behavior of the LDZ with the variation of external and/or internal conditions of a system. The main point of interest of the foregoing analysis is the possibility of computing and forecasting a system response (namely the time course of the informational entropy) to a change of the being conditions. Denote values of the quantities, corresponding to a new set of conditions letters with an asterisk. The change of conditions may lead to both the modification of the regime parameter α^* and the probabilities P_0^* and P_1^*. Timing of time t^* started again at the moment of variation of the conditions. Solution of the differential Kolmogorov equations for the probabilities $P_0^*(t^*)$ and $P_1^*(t^*)$ was obtained [11].

An important role is played by analysis of the functions of entropy flow $H^*(t^*)$ and its rate dH^*/dt^* as a response of an open system to a sudden change of the external and/or internal conditions. Using Eq. (8), a mathematical expression for the time course of the informational entropy flow $H^*(t^*)$ under new conditions was obtained in following form:

$$H^*(t^*) = -\frac{1}{\ln 2(1+\alpha)(1+\alpha^*)}\left\{\left[(1+\alpha)+(\alpha^*-\alpha)e^{-\beta^*t^*}\right]\ln\frac{(1+\alpha)+(\alpha^*-\alpha)e^{-\beta^*t^*}}{(1+\alpha)(1+\alpha^*)} + \right.$$
$$\left. +\left[\alpha^*(1+\alpha)+(\alpha-\alpha^*)e^{-\beta^*t^*}\right]\ln\frac{\alpha^*(1+\alpha)+(\alpha-\alpha^*)e^{-\beta^*t^*}}{(1+\alpha)(1+\alpha^*)}\right\}.$$ (12)

A function of the informational entropy flow rate after a variation of conditions of the system existence can be expressed as:

$$\frac{dH*}{dt*} = \frac{v\mu*-v*\mu}{\beta\ln2}e^{-\beta*t*}\ln\left[\frac{v*\beta+(v\mu*-v*\mu)e^{-\beta*t*}}{\mu*\beta+(v*\mu-v\mu*)e^{-\beta*t*}}\right]. \tag{13}$$

The initial rate of the entropy flow is equal to

$$\left.\frac{dH*(t*)}{dt*}\right|_{t*=0} = \frac{v\mu*-v*\mu}{\beta\ln2}\ln\alpha. \tag{14}$$

It is evident that the obtained Eq. (12) to Eq. (14) depend on the combination of the previous and new being conditions.

Figure 6 shows the graphs of the informational entropy flow $H*$ and its rate $dH*/dt*$ for the case when being conditions become heavier ($\alpha<1$, $\alpha*>1$). On the first time interval ($0<t<1.4$ sec) when $\alpha<1$, the LDZ exists in an easy being mode, so the entropy flow is stabilized at a level H_{st}, corresponding to this mode. At the time $t = 1.4$ sec, being conditions become heavier ($\alpha*>1$). An open system responds to a strong change of conditions by a rapid increase of the entropy flow from the stationary level H_{st}, obtained under previous being conditions, to the maximum value $H*(t_b*) = 1$. At the time t_b*, the rate of entropy flow sharply falls up to zero. The mathematical expression for the moment of time t_b* was obtained in the following form:

$$t_b^* = -\frac{1}{\beta*}\ln\frac{(\alpha*-1)(\alpha+1)}{2(\alpha*-\alpha)}. \tag{15}$$

One can note that it depends on a combination of the previous and new being conditions. In a time t_b* after transition in a heavier being mode ($\alpha*>1$), the LDZ passes across a bifurcation point. This point is associated with destruction of the pattern of the first hierarchical level, exhausting its dissipative possibilities, and emergence of a new appropriate pattern, corresponding to the changed being conditions. Leaving on the second level of hierarchy after the jump of development, the LDZ enters an evolutionary stage of development. There is the rather slow stabilization of the entropy flow during this stage at the expense of saturation by the information up to a level $I = 1- H_{st}*$, which corresponds to a new mode of system existence. In other words, the LDZ adapts to new being conditions by perfection of structure. In this case, the entropy flow rate (See Fig. 6) gets a negative value, passes through a minimum and, remaining negative, aspires to zero, when the transient process reaches the stationary stage with a new steady level of the entropy flow $H_{st}*$:

$$H_{st}* = \ln\left[(1+\alpha*)\Big/(\alpha*)^{\frac{\alpha*}{1+\alpha*}}\right]\Big/\ln 2, \tag{16}$$

adequate to new conditions.

In passage through the critical point at time t_b* the variance $D*_X$ of the systems determining parameter X is maximum. The system is characterized at this stage by the highest degree of disordering, with the random fluctuations manifested on the macroscopic level. After passage through the critical point, the variance $D*_X$ is stabilized on a new stationary level corresponding to the changed being conditions.
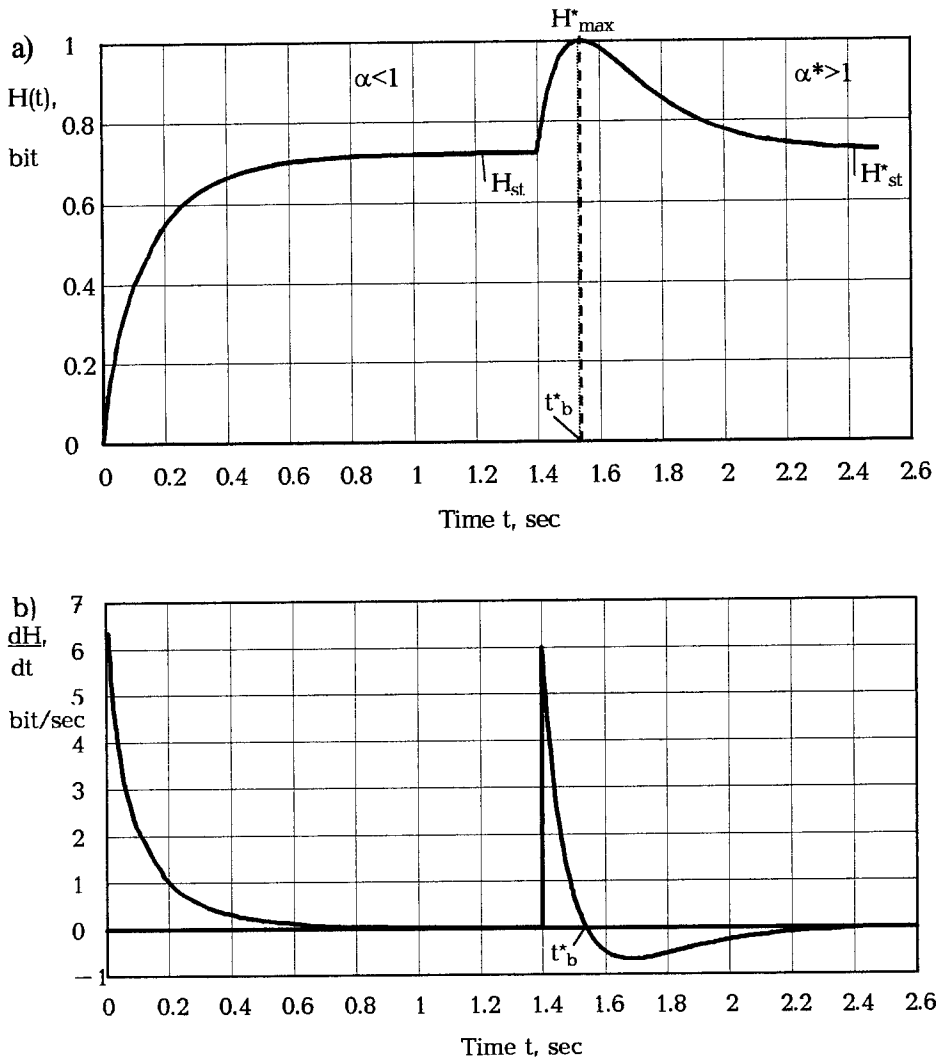
Figure 6. Response of a system to a change of the being conditions. The time courses of the informational entropy flow H(t) (a) and its rate dH/dt (b) for the case when mode of being becomes heavier ($\alpha<1$, $\alpha^*>1$). The coordinates of the critical point are labeled as H^*_{max} and t^*_b.

There is no place for more detailed examination of behavior in the course of time of the statistic properties of the systems determining parameter X after changing the being conditions. Let us note only, that there are possible both the cases of monotonous increase or decrease of the variance D^*_X and the cases of emerging the characteristic peak on a curve of variance D^*_X. The characteristic peak is similar to that shown in Figure 3 and corresponds to a critical point. Passing through such a typical peak and stabilizing the variance D^*_X on a new steady level D^*_{ST} confirm completion of the period of running in ("burn-in") of a system and its adaptation to new changed conditions. When testing the system, it is possible to measure and record the time history of the mean value and

variance of determining parameter in the vicinity of some LDZs. We believe that on a basis of analysis of processes, occurring in the LDZ, it should be possible to speed up the reception of data concerning the termination of the transition to new state as well as adaptation of system to changed conditions. In Sec. 6 we shall consider the plots, representing the time evolution of the mean value and the variance for process of nucleation in model system involving multiple steady states [8,12].

An open system of any nature comes back in a steady stable state due to the inflow of information from the outside or/and redistribution of the informational entropy between hierarchical levels of the system [2].

After passage through the bifurcation point, the entropy flow decreases in accordance with an information accumulation; that means appropriate increase of an organization level during a system development. At each hierarchical level of a system's evolution at the end of self-organization process, when the "architecture" of the system was basically defined and becomes saturated by the information, the entropy curve is gradually straightened, displaying the transition of a system to the evolutionary stage of development. A growth in the degree of organization for any system has a limit, area of saturation, determined by the limited opportunities of an information accumulation in the given structure at a given hierarchical level.

Such a picture of temporal evolution of an open system with respect to variation of the being conditions agrees with the synergetic approach to processes of self-organization in nonequilibrium dissipative systems [1,4,8,9]. Obtained analytical dependencies reveal a quantitative ratio, reflecting the dynamics of the entropy flow and its rate, which shows the evolution of open systems in the course of its life cycle under variation of being conditions. The derived laws, in our opinion, are applicable for open nonequilibrium systems of various natures: technical, economic, biological, ecological, social, etc. The discussed model and obtained dependencies make it possible on a unified basis to describe the whole life cycle of a system, including passage across a sequence of bifurcation points ("jumps" of development) and evolutionary stages of development at each hierarchical level. Transition to a new level of development goes from disorder to order, through the phenomenon of instability in the bifurcation points, where a system has a number of options to diverge in several directions.

In the following, we shall give a few examples in order to elucidate the received results.

5 Example 1. Self-Organization during Elasto-Plastic Deformation of Metal

Here we consider an example of self-organization process during elasto-plastic deformation of metal in local zones of load-carrying structures. From the thermodynamic and synergetic points of view, a material undergoing plastic deformation is an open system brought far from equilibrium conditions [4,6,10]. In this case, the LDZ is represented by the zone of stress (strain) concentration. Let us accept as determining parameter X, limiting the lifetime of the load-carrying structure, stress intensity σ_i, determined from the distortion energy (Huber–Von Mises–Hencky) theory. It was assumed that transition from down state, corresponding to the elastic behavior of metal in the LDZ, in up (the plastic yielding state) is caused by such an external influence, exerted upon the system, under which the determining parameter exceeds a threshold level equal to the mean value of the yield strength σ_s, that is: $\sigma_i \geq \sigma_s$.

The main point of this example is to show how the random loading process exerted upon the load-carrying structure is linked with the structural response (namely the time

course of entropy) to a change of the loading conditions. Let us accept the hypothesis that instantaneous values of the loading process have a normal distribution with the following statistical properties: mean value σ_m and variance D_σ of the stresses. Using Eq. (6), the relations that link the regime parameter α with these statistical properties and also with the threshold level (mean value of yield stress σ_S) were obtained in the following form:

$$\alpha = \left[\Phi\left(\frac{\sigma_S - \sigma_m}{D_\sigma} \right) \right]^{-1} - 1 \qquad \text{at } \sigma_m < \sigma_S , \tag{17}$$

$$\alpha = \left[1 - \Phi\left(\frac{\sigma_m - \sigma_S}{D_\sigma} \right) \right]^{-1} - 1 \qquad \text{at } \sigma_m > \sigma_S , \tag{18}$$

where

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} \exp\left(\frac{-x^2}{2} \right) dx, \, Z = \frac{\sigma_S - \sigma_m}{D_\sigma} . \tag{19}$$

The dependence of α on σ_m and D_σ for low-carbon steel was computed from Eq. (17), (18) and plotted in Figure 7. We used this approach for a load-carrying welded structure, namely the pivoting section of a gondola car body. Examination of the time history of stress in elements of the gondola car, including directly the welded joints, carried out using the automated system for amplitude-spectral analysis, shows that the PDF of the instantaneous values of the random process is normal.

In the course of elasto-plastic deformation in metals, a number of dislocation patterns (microstructures) are formed. One after the other ball, cellular, persistent slip band, quasi-amorphous microstructures arise and are destroyed, consistently replacing each other on a background of existing grains' boundaries. Further increase of load results in formation of the crack origins in a quasi-amorphous zone and growth of their density. The spontaneous emergence of a quasi-amorphous microstructure corresponds to achievement of a maximum disorder in this local zone, at which point the thermodynamic entropy is maximum and equal to the enthalpy of melting. All these transitions are supervised by achievement of a maximum level of the entropy flow [4]. Under action of an energy flux, pumped up by the stochastic loading process in the LDZ, the deformation ability of metal at the lowest hierarchical structural level is exhausted. Getting through a critical point of bifurcation, metal passes to a higher level of the pattern hierarchy to microstructure, having the higher dissipative characteristic [4]. This transition between patterns on the microscopic level is reflected by the curves of the entropy flow and its rate, obtained before. Returning to Figure 6, note that at the first time stage ($0 < t < 1.4$ sec) the regime parameter α was determined by using Eq. (17) to be equal to 0.25. It corresponds to the vertical (gross) dynamic load of the gondola car. The value $\alpha^* = 5$ corresponds to the total effect of the vertical and longitudinal inertial (in collision of cars) loads. That mode of the LDZ being matches the transition of metal in plastic state. In Figure 6, one can see that a load-carrying structure responds to the heavy conditions by a rapid increase of the entropy flow from the stationary level H_{st}, obtained under vertical load, to the maximum H^*_{max} at the time t_b^*. This critical point is associated with the destruction of the pattern exhausting its dissipative abilities, and the emergence of new microstructure at another hierarchical level corresponding to plastic state of the LDZ. The local zone becomes structured, responding to the heavy being mode by self-organization of more advanced patterns.

240

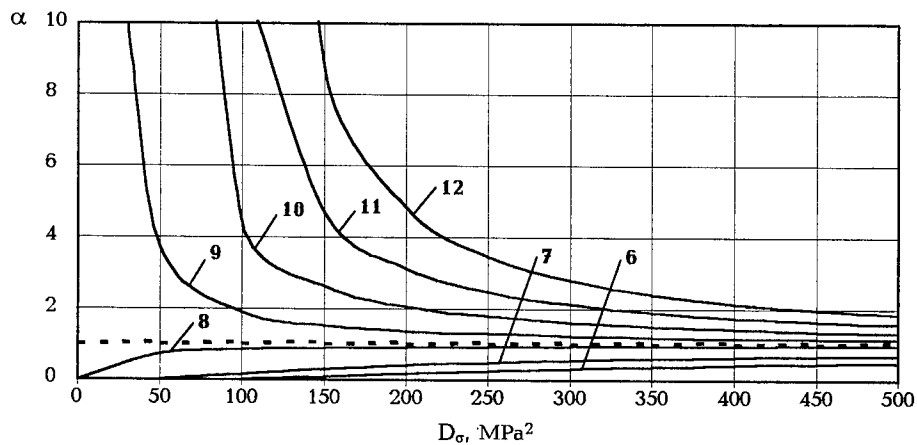a)
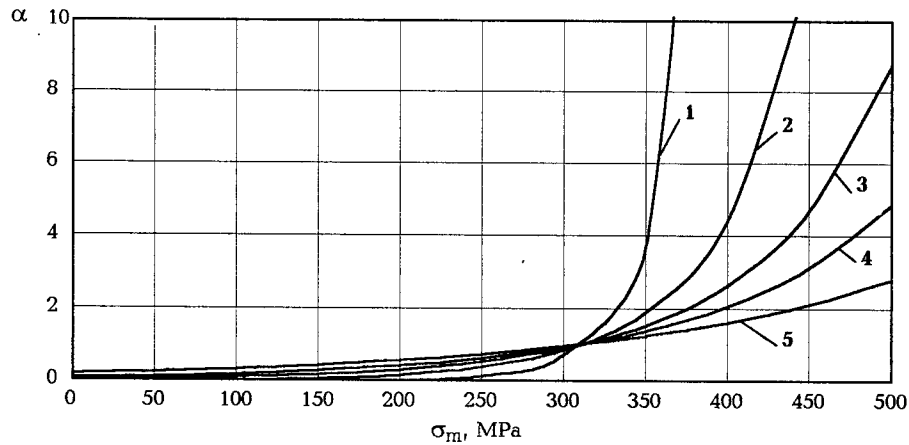
b)

Figure 7. Change of the regime parameter α, as a function of the statistical properties (mean value σ_m (a) and variance D_σ (b)) of Gaussian random process of dynamic stresses: 1 − 5) D_σ is equal to 50, 100, 150, 200, 300 MPa^2 respectively; 6 − 12) σ_m is equal to 100, 200, 300, 350, 400, 450, 500 MPa respectively. Line $\alpha = 1$, corresponding to threshold level $\sigma_m = \sigma_s = 310$ MPa, is denoted by a dashed line.

6 Example 2. Nicolis and Prigogine's Model

A process of nucleation caused by formation of germs in model system involving multiple steady states is considered by Nicolis and Prigogine [8,12]. The plots, representing the time evolution of the average value and the variance for this numerical model, are indicated in Figure 8, which was adopted from [12].
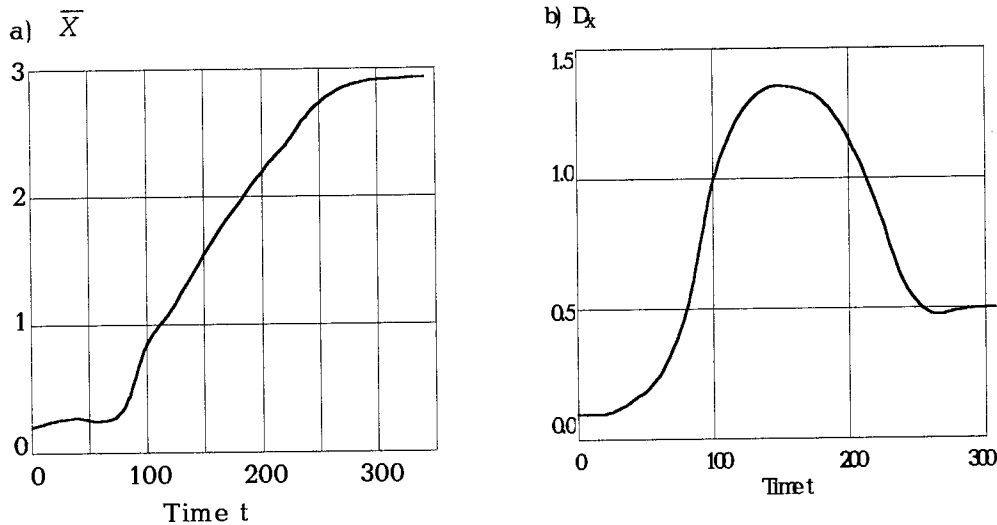
Figure 8. Evolution of the mean value $\overline{X}$ (a) and variance D_X (b) of fluctuations for the Nicolis-Prigogine's model. Adopted from [12].

We can see the process of the model system transition from initial state to another stable state corresponding to a higher level of mean value. During the process the mean value of determining parameter monotonously increases (See Fig. 8a). The variance begins to increase sharply as soon as sufficiently large new domain appears in the system through fluctuations. Appearance of the typical peak on the curve $D_X(t)$ and subsequent stabilization of the variance on a new steady level (See Fig. 8b) lead to a conclusion that the transition period is finished.

The plots of Figure 8 correspond to analytical dependencies, Eqs. (2), (3) of mean value $\overline{X}(t)$ and variance $D_X(t)$ of determining parameter X with time (See Figures 2 and 3). We believe that on a basis of the analysis of processes, occurring in the LDZ, it should be possible to speed up the reception of data concerning the termination of the transition as well as adaptation of system to new conditions.

7 Conclusions

The various kinds of open systems (technical, ecological, biological, socio-economic, etc.) respond to a strong change of being conditions in the same way: by steep growth of the entropy flow up to a maximum value at the critical point. According to the traditional interpretation of entropy, it means that the disordering and chaos in the system increase at this stage of its evolution. The critical point (being the stochastic analogue of a bifurcation point) associates with the self-organization process, that is the destruction of pattern of the previous hierarchical level, exhausting its possibilities, and resulting in the emergence of new more complicated and more advanced patterns, corresponding to the changed being conditions. An open system of any nature comes back to a steady stable state due to inflow of the information from the outside or/and redistribution of the informational entropy between the hierarchical levels of system.

Obtained mathematical expressions for the time course of the informational entropy flow make it possible to predict the moment that the critical point will be reached. That is the moment of maximum uncertainty, instability, and chaos in the system when small fluctuations become amplified up to a macroscopic scale. It may drive the open system into a completely new state and thus become the driving force of the system's development.

Acknowledgements

REFERENCES

1. Alonso M. (ed.), *Organization and Change in Complex Systems* (Paragon House, New York, 1990).
2. Auger P., *Dynamics and Thermodynamics in Hierarchically Organized Systems: Applications in Physics, Biology and Economics* (Pergamon, Oxford, 1989).
3. Ingarden R. S., Kossakowsky A. and Ohya M., *Information Dynamics and Open Systems: Classical and Quantum Approach* (Kluwer Academic Publishers, Dordrecht, 1997).
4. Ivanova V. S., *Synergetics: Strength and Fracture of Metallic Materials* (Cambridge University Press, 1998).
5. Krohn W., Kuppers G. and Nowotny H. (eds.), *Selforganization: Portrait of a Scientific Revolution* (Kluwer Academic Publishers, Dordrecht, 1990).
6. Kubin L. P. and Lepinoux J., The Dynamic Organization of Dislocation Structures, In *Proc. of VIII Intern. Conf. of Strength Metals and Alloys* (Tampere, 1988) **1** pp. 35-59,
7. Nicolis J. S., *Dynamics of Hierarchical Systems: An Evolutionary Approach* (Springer, Berlin, 1986).
8. Nicolis G. and Prigogine I., *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations* (Wiley, New York, 1977).
9. Prigogine I., *From Being to Becoming: Time and Complexity in the Physical Sciences* (W.H.Freeman and Company, New York, 1980).
10. Seeger A., Thermodynamics of Open Systems, Self-organization, and Crystal Plasticity, In *Proc. of VIII Intern. Conf. of Strength Metals and Alloys* (Tampere, 1988) **1** pp. 463-468.
11. Zainetdinov R. I., Dynamics of Informational Entropy Associated with Self-Organisation Process in Open System, *Chaos, Solitons & Fractals* **9** (1999) pp. 1425-1435.
12. Николис Г., Пригожин И. *Познание сложного: Введение.* Пер. с англ. (Мир, Москва, 1990).

FRACTAL PROPERTIES IN ECONOMICS

HIDEKI TAKAYASU

Sony Computer Science Laboratories Inc.,3-14-13, Higashigotanda, Shinagawa-ku,
Tokyo, 141-0022, Japan
E-mail:takayasu@csl.sony.co.jp

MISAKO TAKAYASU
Faculty of Science and Technology, Keio University,
890-12 Kashimada, Saiwai-ku, Kawasaki 211-0958, Japan
E-mail:misako@future.st.keio.ac.jp

MITSUHIRO P. OKAZAKI
Department of Industrial Engineering, Musashi Institute of Technology,
1-28-1, Tamazutsumi, Setagaya-ku, Tokyo, 158-8557, Japan
E-mail:f6586719@hi-ho.ne.jp

KOUHEI MARUMO
Institute for monetary and economic studies, Bank of Japan
2-1-1 Hongoku-cho Nihonbashi, chuo-ku, Tokyo, 103-8660, Japan
E-mail:kouhei.marumo@boj.or.jp

TOKIKO SHIMIZU
Financial markets department, Bank of Japan
2-1-1 Hongoku-cho Nihonbashi, chuo-ku, Tokyo, 103-8660, Japan
E-mail:tokiko.shimizu@boj.or.jp

Scaling properties in financial fluctuations are reviewed from the standpoint of statistical physics. We firstly show theoretically that the balance of demand and supply enhances fluctuations due to the underlying phase transition mechanism. By analyzing tick data of yen-dollar exchange rates we confirm two fractal properties: 1 The distribution of rate change in a fixed ticks is approximated by a symmetric stretched exponential function for a wide range of time intervals; 2 the interval time distribution of trades nearly follows a power law. Empirical fractal properties in companies' financial data, such as distributions and fluctuations in assets and incomes are discussed with a simple model. The importance of methods and theories for phase transitions is discussed.

1 Introduction

The concept of fractals has been spread over all fields of sciences and technology, however, it seems less known that this concept was born in the field of economics when Mandelbrot was investigating the price changes in an open market around 1963 [1]. He found empirically that a chart of market price changes of cotton price looks similar to another chart with different time resolution. A general concept flashed across his mind that such scale invariance could be the clue to characterize many complex phenomena around us. After about 20 years the concept he created, the fractals, became a key word in a wide research activity. Now, more than 35 years have passed since his original discovery, economics becomes one of the hottest topics in the study of fractals.

A new field of science, econophysics, was established recently in 1997 [2, 3]. This is a study of economic phenomena based on the methods and approaches of physics. Among

many topics in econophysics there are three topics that are closely related to the study of fractals. They are price changes in open market, the distribution of income of companies, and the scaling relation of company's size fluctuations. In this paper we briefly review these topics and discuss the validity of the known empirical laws introducing simple mathematical models for better understanding.

In the following Section we discuss price fluctuations which naturally involve critical fluctuations near the equilibrium point of demand and supply. In Sec. 3 we show results of tick data analysis of foreign exchange rates. We show theoretically that the fat tails of the rate change distribution can be modeled by a very simple physical equation of Langevin with random coefficient. In the Sec. 4 we review the empirical laws on companies' macroscopic financial data of annual income and assets. We introduce a very simple model of company growth that reproduces the Zipf's law in income distribution, also we discuss a nonlinear relation between income and assets. Sec. 5 is devoted for general discussion and concluding remarks on the importance of statistical physics for the data analysis in economics.

2 Demand and supply; a phase transition view

In economics, it had been known for a long time that demand and supply balance automatically. However, it becomes evident that in reality such balancing is hardly realized for most of popular commodities in our daily life [4, 5]. The important point is that demand is essentially a stochastic variable because human action can never be predicted perfectly, hence the balance of demand and supply should also be viewed in a probabilistic way. If demand and supply are balanced on average, the probability of finding an arbitrarily chosen commodity on the shelves of a store should be 1/2, namely about half of the shelves should be empty. Contrary to this theoretical estimation, shelves in any department store or supermarket are nearly always full of commodities. This clearly demonstrates that supply is much in excess in such stores. Excess supply generally holds for most of commodities and especially for food in economically advanced countries. For example, it is reported in a newspaper that about a quarter of food produced in the USA is disposed as garbage. If food supply were equal to demand on average then it implies that about half of the population were starving!

In general, the stochastic properties of demand and supply can be well characterized by a phase transition view, which consists of two phases: the excess-demand and excess-supply phases. It has been mathematically proven that for commodities whose production costs are smaller than a certain portion of the selling prices, an excess-supply strategy becomes the best strategy, when the demand fluctuates randomly and unsold commodities are to be disposed [6]. Obviously department stores follow this excess-supply strategy. The other strategy of keeping excess-demand can also be a rational solution, when the production cost and the selling price are close. The critical point at which the averaged demand and supply are equal realizes only in a very special case.

It is a general property of a phase transition system that fluctuations are largest at the phase transition point and this property holds also in this demand-supply system. In the case of markets of ordinary commodities, consumers and providers are independent and the averaged supply and demand are generally not equal. The resulting price fluctuations are generally slow and small in such market because the system is out of the critical point. On the contrary in an open market of stocks or foreign exchanges, market is governed by speculative dealers who frequently change their positions between buyers and sellers. It is shown that such speculative actions make demand and supply balance automatically on

average by changing the market price [5]. As the system is always at the critical point, the resulting price fluctuations are generally quick and large involving fractal properties.

3   Price changes in open markets

The physical mechanism of market price fluctuations was firstly studied by one of the authors (H. T.) et al in '92 by introducing a numerical market model[7]. The model consists of speculative dealers who transact with others simply following the basic rule "buy at a lower price and sell at a higher price". These two threshold prices are determined at each time step by each dealer taking into account the information of past market price changes. It is shown that even a smallest limit case of 3 dealers can show chaotic behavior, implying that the transaction's nonlinear effect is very strong. The resulting market prices naturally include sudden falls or increases when the parameter controlling the dealers' averaged response to latest price change is large enough.

In '95 a famous paper by Mantegna and Stanley appeared in Nature in which price changes in a stock market are reported to follow a power law distribution with truncation at higher values [8]. The exponent of the power law is estimated as 1.4 in the form of stable distribution's characteristic exponent. However, this power law has not been accepted widely and the precise functional form of the distribution of open market price changes is still under intensive debate.

Here, we show our original results of analysis on tick data of yen-dollar exchange rates for about six months. Tick data are the highest resolution transaction data of foreign exchange markets, which are announced to the dealers electronically. Transactions of foreign exchanges are done by telephone calls, so the market is active when more than 2 dealers are working on the globe, namely, continuous tick data can be obtained except for weekends. The tick data monitor several representative dealers' transactions and the number of data is about 12,000 every day. We analyzed a half year data from October '98 to March '99, and the total number of ticks exceeds a million.

Fig.1 shows a typical example of yen-dollar rate changes in 3 different time scales. Intuitively, this figure demonstrates a fractal property of exchange rates in the time axis measured by ticks, namely, Mandelbrot's classical finding also holds for this contemporary market price fluctuation. The statistics of this fluctuation is very close to random walk, actually it is easy to confirm that the power spectrum of this fluctuation clearly follows an inverse square law, that is almost identical to a Brownian motion. The corresponding auto-correlation function for rate fluctuation per tick defined by the following equation decays very quickly as shown in Fig.2.

$$C(T) = \frac{<\Delta r(T_0 + T)\Delta r(T_0)> - <\Delta r(T_0)>^2}{<\Delta r(T_0)^2> - <\Delta r(T_0)>^2} \tag{1}$$

where $\Delta r(T)$ denotes the rate change at T-th tick, namely, $\Delta r(T) = r(T) - r(T-1)$ with $r(T)$ being the exchange rate at T-th tick time, and $<...>$ shows average over tick times T_0. The correlation is virtually 0 even at $T=2$, which corresponds to the averaged physical time of only 15 seconds. As seen from this property, the rate change fluctuation is very close to a white noise.

The distribution of rate change in a tick is shown in Fig.3 on semi-log plot with a best-fitted normal distribution for comparison. The tails are much fatter than the normal
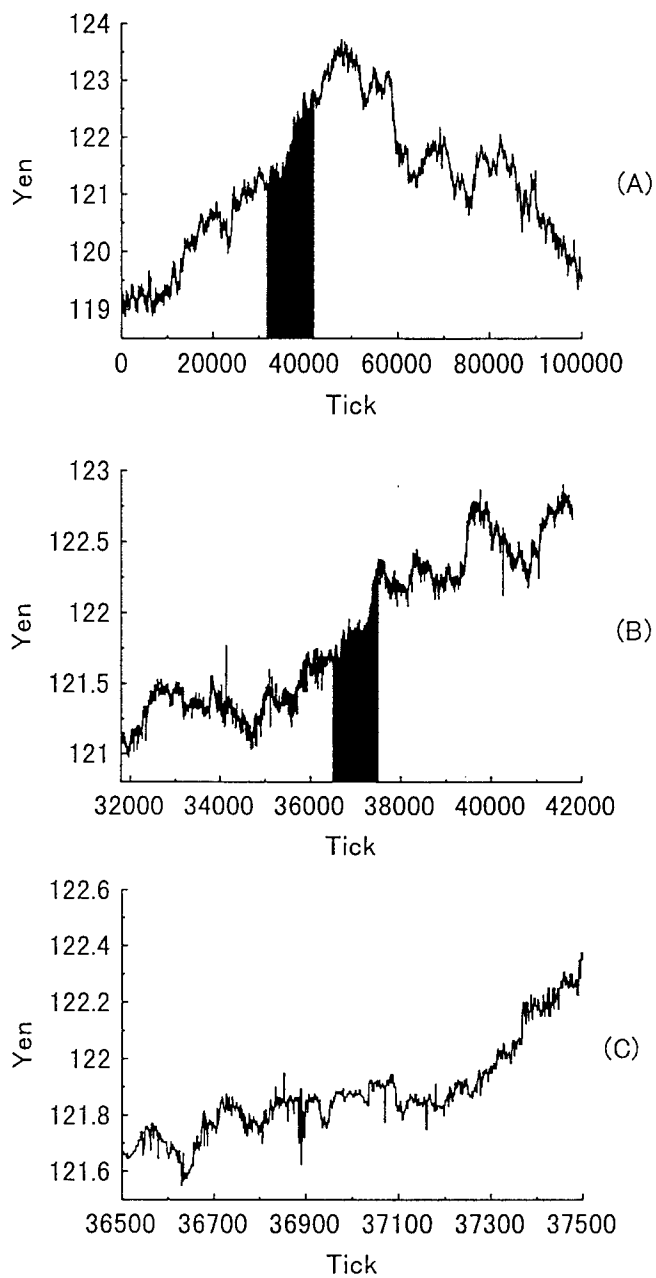
Figure. 1 Fractal property in foreign currency exchange rate fluctuations.
The dark part of (a) is magnified 10 times in (b), and the dark part in (b) is magnified in (c).
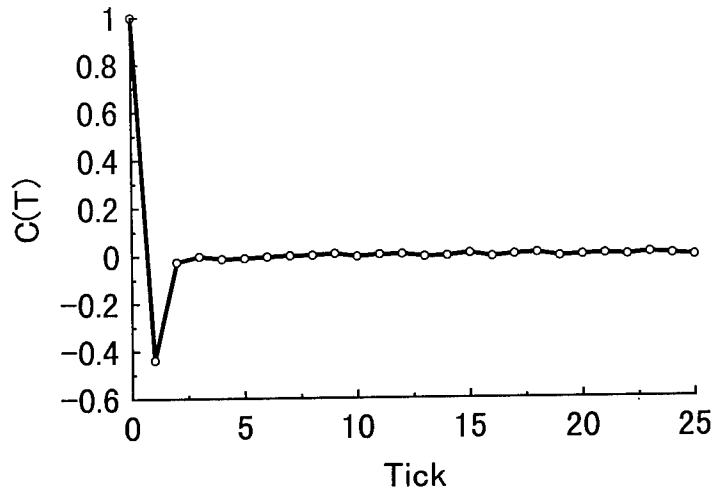
Figure. 2 Autocorrelation of rate changes observed by tick data.

distribution and we can find extreme events deviating from the mean value, $< \Delta r > \doteq 0.0$, with magnitude nearly 50 times larger than the standard deviation. In order to clarify the functional form of this distribution we plot a cumulative distribution in log-log scale in Fig.4. The distribution for Δr in the negative part is plotted by taking absolute values. The distribution functions for both tails are quite identical except for the very large events for which the number of data is limited. The tails decay gradually in log-log scale and we can not find a linear slope. We can approximate the whole shape of the distribution by a symmetric stretched exponential function as

$$P(\geq| \Delta r |) = 0.28\exp[-7.0(| \Delta r |-0.04)^{0.4}] \quad , \quad | \Delta r | \geq 0.05 \qquad (2)$$

The dashed curve gives this function which is plotted with a shift for better comparison. For small absolute values the function does not fit with this analytic form because real data take discrete values which are given by 0.01 times integers. Omitting about 10 points around the center the theoretical fitting is quite plausible in a wide range. Note that this function decays gradually in the log-log plot, so if we dare to fit a linear slope then we might deduce a non-universal power law that might depend on the fitting range.

We observe daily change of the distributions of rate fluctuations per tick and find that the distributions are always quite symmetric and the functional forms are always similar to Eq.(2), however, the standard deviation can change more than ten times from 0.03 upto 0.36. This phenomenon is related to the problem of long time correlation in volatility fluctuations observed in any open market price [3]. We can easily confirm that volatility measured by the absolute values of the rate changes has a long correlation, however, the details of its functional forms are yet to be clarified.
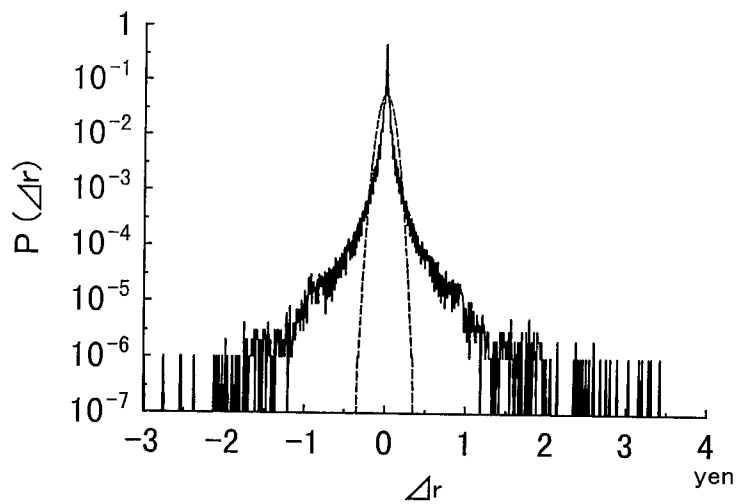
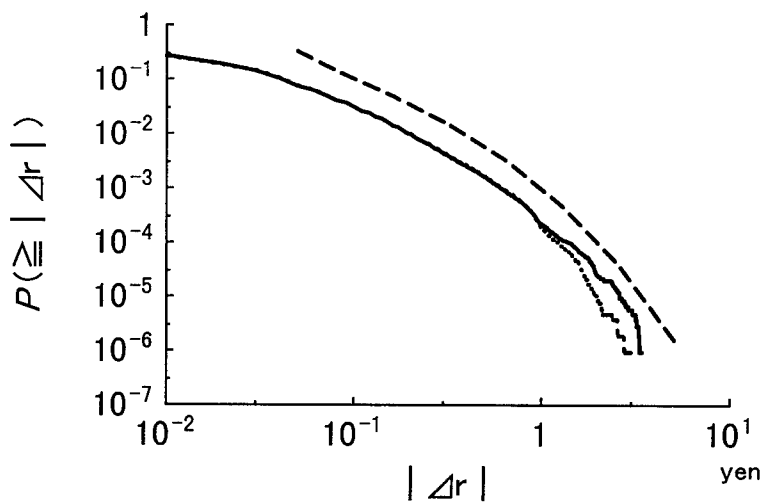Figure. 3 Rate change distribution viewed by tick data in semi-log plot.

Figure. 4 Log-log plot of the cumulative distribution of rate changes. Positive tail (the bold line), negative tail (dotted line) and the theoretical curve fitted by a stretched exponential function, Eq.(2) (dashed line).
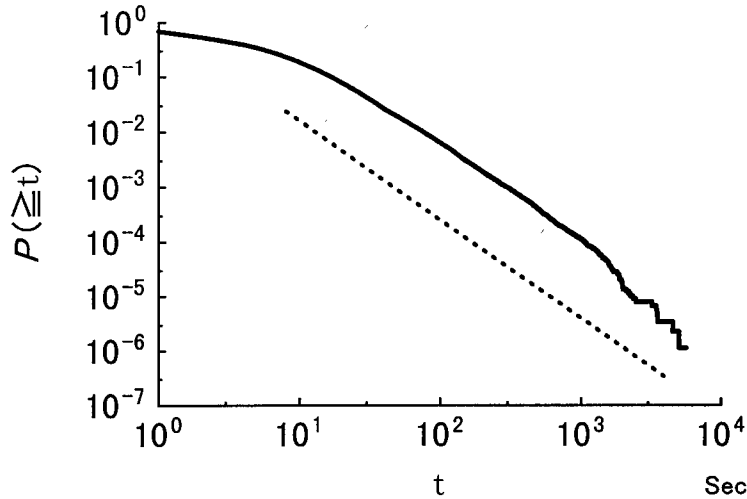
Figure. 5 Distribution of time interval between two successive trades in log-log scale.
The dotted line shows a power law with exponent −1.8.

The time measured by tick is obviously different from the physical time. The occurrence of trades is also a stochastic process and we observe the distribution of tick intervals. In Fig.5 the cumulative distribution of tick intervals measured by seconds are plotted in log-log scale. Here, the minimal tick interval is 0 second and the largest is about 5600 seconds, and intervals longer than this are neglected because they include weekends or holidays. The scale range is therefore limited, however, we can find a power law tail in the scale range of [30, 1600] in seconds.

These statistical properties are not limited to yen-dollar rates but they are applicable to other major currency exchanges. Similarities can be extended to stock markets or other open markets to some extent, however, in such markets trading volumes are much smaller than currency exchange. The total money flow in foreign currency market in one day is estimated to be more than $2*10^{14}$ yen which is roughly a half of the annual gross domestic product (GDP) of Japan. Other than foreign exchanges the markets are open only in office hours, so the statistics are expected to be less universal than the case of currency markets. For example, currency market fluctuations are generally symmetric, however, in stock markets there sometimes appear asymmetric properties such as crashes. We believe that universal properties can be more clearly observed in the foreign currency exchange markets than any other markets.

The fat tail properties in the rate changes can be explained by introducing a simple mathematical model. It is deduced from the analysis of numerical market model that the market price changes can be approximately described by a Langevin equation with fluctuating coefficient [9].

$$\Delta r(t + \Delta t) = b(t)\Delta r(t) + f(t) \tag{3}$$

Here, in this formulation the time t can be either tick time or physical time, $f(t)$ and $b(t)$ represent the random force term and random coefficient, respectively. The random force term comes from the scatter of dealers' characters and predictions. The value of $b(t)$ is larger when the dealer's averaged responses to the latest price change is larger. It is known that statistical properties of $\Delta r(t)$ are fully characterized by the statistics of $b(t)$ [10]. In the case that $b(t)$ is always smaller than 1 the distribution of $\Delta r(t)$ is known to follow a stretched exponential form. If the probability of $b(t)$ taking a value larger than 1 is not zero, the distribution of $\Delta r(t)$ has power law tails. The exponent of the power law, β, is given by a simple formula,

$$P(\geq | \Delta r |) \propto | \Delta r |^{-\beta} \quad \text{with} \quad < b(t)^{\beta} > = 1 \tag{4}$$

here $<...>$ denotes ensemble average. Eq.(4) is mathematically rigorous in the range $0 < \beta < 2$. When large values appear successively in $b(t)$, Δr is enhanced successively by the recursion of Eq.(3) resulting a large value in Δr.

An interesting point in this discussion is that Langevin equation is very popular in a wide area of physical systems, therefore, a situation that is similar to the case of open market can be realized in purely physical systems. For example, one of the authors (H.T.) and his collaborators are now developing an electrical circuit that has a randomly fluctuating resistance. Note that the case $b(t)>1$ corresponds to a negative resistance, that is, an amplification, in the continuum limit of Eq.(3). According to our preliminary results the statistics of its voltage fluctuations caused by thermal noise shows similarity with that of price changes or rate changes in open markets [11].

The model equation, Eq.(3), captures a basic aspect of price changes in open market, however, it is obviously too simple. Two of the authors (H.T. and M.T.) have already derived a more general basic equation of price changes by theoretical considerations of the dynamics of dealers [5]. Skipping the derivation, the general equation is a set of linear equation with random force and random coefficients as

$$r(t + \Delta t) = r(t) + a(t)(r^*(t) - r(t)) \tag{5.a}$$
$$r^*(t + \Delta t) = r^*(t) + f(t) + b(t)(r(t) - r(t - \Delta t)) \tag{5.b}$$

where $r^*(t)$ is a virtual equilibrium price that is determined by the balanced price of demand and supply when all dealers show their buying and selling prices in mind openly. The coefficient $a(t)$ shows the market price's response against the change of demand and supply that is inversely proportional to the price rigidity in economics, namely, for larger $a(t)$ the market price changes more largely for the same modification of demand-supply configuration. For given $\{a(t)\}$ and $\{b(t)\}$ with initial condition for $r(t)$ and $r^*(t)$, Eq.(5a) and (5b) solve the time evolution of $r(t)$ and $r^*(t)$ simultaneously. In the special limit that $r(t)$ is always nearly equal to $r^*(t)$ these set of equations become identical to Eq.(3), then the basic properties of Eq.(3) are recovered.

There are 3 typical behaviors in the time evolution of Eqs.(5). One is a stable state in which the deviation between $r(t)$ and $r^*(t)$ does not increase on average which is realized when both $a(t)$ and $b(t)$ are smaller than 1. Another behavior is an exponential growth of $r(t)$ which corresponds to the phenomenon called "bubble". This behavior occurs when $b(t)$ is larger than 1 and $a(t)$ is smaller than 1, namely when the dealers are strongly

expecting the latest trend to be continued and the price rigidity is rather high. In such case the expected equilibrium price $r^*(t)$ goes ahead while $r(t)$ follows it, and the deviation between $r(t)$ and $r^*(t)$ increases exponentially. The third behavior is an oscillatory fluctuation that happens when $a(t)$ is larger than 1. In such case the price rigidity is so small that the market price goes over $r^*(t)$ causing an oscillation.

There are two obvious defects in the simplest model equation, Eq.(3); the lack of long time correlation in volatility and the absence of characteristic behaviors such as bubbles and oscillations which are observable in real open market data. We are now studying the basic properties of the generalized version, Eqs.(5), comparing with real data, so that the values of the coefficients, $a(t)$ and $b(t)$, can be estimated from the real time sequences of $r(t)$.

4   Company's macroscopic financial data analysis

History of study on fractal properties in money flow can be traced back more than 100 years. In 1896 Pareto investigated personal incomes in Italy and found that the distributions are approximated by a power law [12]. In 1922 Gini explored the same quantity for several European countries and reported that the exponents of the power laws are different for different country [13]. Two physisists, Montroll and Schlesinger, paid attention to the power law nature of income distribution in 1983 and found that personal income of top 1% follows a power law while the rest of 99%, who are mostly salaried, are characterized by a log-normal law [14]. Recently, fractal properties of company's income and asset are attracting much interest.

M. H. R. Stanley et al established an interesting scaling relation on the statistics of growth of companies [15]. Let $A(t)$ be the asset of a company at the t-th year, where asset means intuitively the price of the company as a whole. Asset growth is characterized by the growth rate that is defined by logarithm of the ratio of asset change,

$$R(t) = \log \frac{A(t)}{A(t-1)} \tag{6}$$

The distribution function of $R(t)$ has been observed for more than 10,000 companies in USA and for about ten years. It is shown that the probability density of $R(t)$ for a company with asset A is approximately given by an exponential function

$$p(R \mid A) = \frac{1}{\sqrt{2}\sigma(A)} \exp(-\frac{\sqrt{2}\,|R - <R(A)>|}{\sigma(A)}) \tag{7}$$

where $\sigma(A)$ is the standard deviation for asset size around $A(t)=A$. There exists a non-trivial scaling relation in $\sigma(A)$ such as

$$\sigma(A) \propto A^{-\theta} \quad , \quad \theta \sim 0.15 \tag{8}$$

Namely, for larger asset companies the fluctuations are relatively smaller, which is intuitively very natural.

252

There are several quantities that characterize the size of companies such as the number of employee and net sales, and it is confirmed that characterization by using these quantities also leads the same result. Also this type of scaling relation is known to hold not only in USA but also for other countries, for example, Japan, France and Italy [16].

If all of the components of a company fluctuate independently then it is easy to show that the value of the exponent θ should be 0.5, while if the growth is merely proportional to the whole asset then θ should be 0. Therefore, the non-trivial value of θ strongly suggests that the company growth is governed by non-trivial interactions either internal origins [17] or external ones [16].

Universality of the scaling relation, Eq.(8), is still wider. It is known to be valid also for the growth of gross domestic product (GDP) of more than 150 countries for the periods of more than 40 years after the world war II [18]. Although the detail mechanisms of these phenomena are yet to be clarified it is likely that gigantic economic systems seem to fall in the same universality class in which non-trivial fractal relation holds.

Very recently two of the authors (H.T. and M.T.) together with Okuyama examined detailed data of annual incomes of companies in Japan which had been reported to the tax office and they found that a clear power law with exponent very close to -1 holds (see Fig.6) [19]. The data cover all Japanese companies whose incomes are larger than $4.0 * 10^7$ yen, and the number of the companies is about 85,000. Here, income is in rough sense given by the total sales subtracted by the total outgoing costs. There are fluctuations for very large incomes, however, for companies having income less than 10^{10} yen the plots are clearly on a straight line in the log-log plot. Such power law behaviors can also be confirmed for other years and for several other countries.
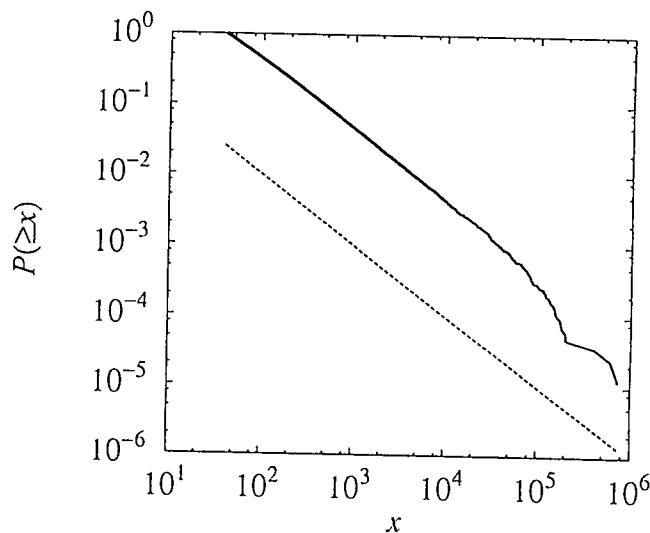

Figure. 6 Log-log plot of the income distribution of Japanese companies for all job categories. Income x is represented in the unit of million yen.
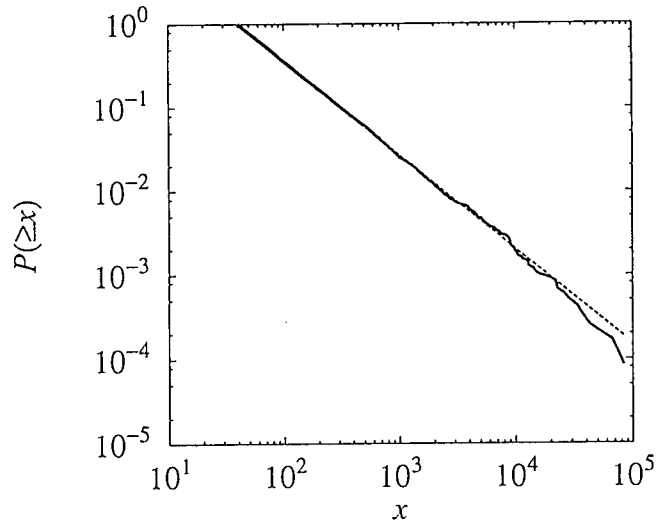
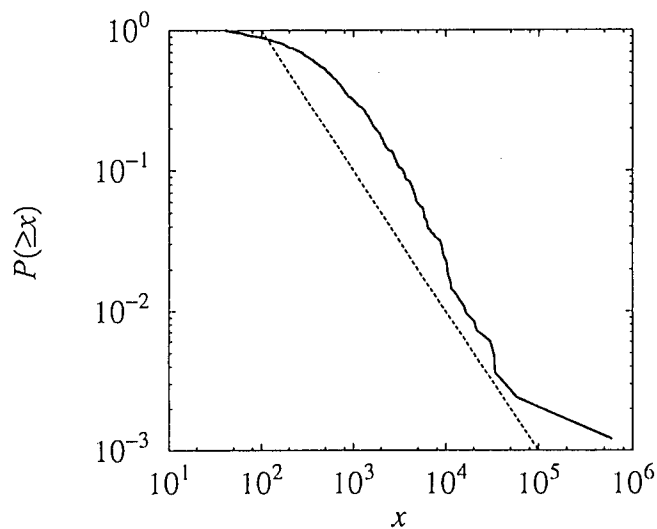Figure. 7 Income distribution for the category "constructions".

Figure. 8 Income distribution for the category "banks".

An interesting feature of this power law distribution is that similar power law holds in each category of industry. For example, Fig.7 shows the same plot as in Fig.6 but the companies are limited only those categorized in "constructions". The number of companies is about 11,000 and the estimated power exponent is -1.1, very close to the case of the whole categories.

By carefully observing the distributions of income in each job category we find that there are some exceptions that do not follow the power law. For example, in Fig.8 we show the distribution of the case of category "banks". Here, we can not find a clear slope in the plot. The existence of bending point around 10^2 million yen implies that the number of banks whose incomes are less than this amount is very little. In Japan there were many regulations by law for banks for example the interest rates were standardized all over the country, which causes little competition among banks, also a new bank could not be established easily. As symbolized in this example deviation from power law distributions tends to be observed in job categories which are less competitive compared with other job categories. Recently, Japanese bank regulations have been loosened drastically, so the future change in the distribution of income of banks should attract not only economists' but also scientists' attention.

What is the relation between asset and income, as both are characterizing company's statistics? To be precise the income defined above includes taxes to be paid to the government that is about 50% independent of the asset or income size in Japan. The rest of the income is normally added to the company's asset, namely, the following relation holds in rough estimation.

$$A(t+1) = A(t) + I(t)/2 \qquad (9)$$

where $I(t)$ denotes the income of the t-th year. It is confirmed from the data that $I(t)$ is generally about 2 orders smaller in magnitude and it can take a negative value, while the company collapses when $A(t+1)$ becomes negative. Comparing Eq.(9) with Eq.(6) we have a relation between $R(t)$ and $I(t)$ as

$$R(t) \propto \frac{I(t)}{A(t)} \qquad (10)$$

The relation between income and asset can be compared directly by plotting these two quantities. Scatters are generally large but the following nonlinear relation has been found as average [19].

$$I(t) \propto A(t)^{\delta} \quad , \quad \delta = 0.85 \qquad (11)$$

These relations (10) and (11) are consistent with Eq.(8) and we confirm a relation between the exponents, $\delta = 1 - \theta$. This nonlinear relation implies that large systems are less efficient, which is intuitively well appreciated.

So far no mathematical model is known that reproduces all of these empirical relations of company's financial data. Here, we introduce a simplest mathematical model based on competitive growth that realizes the power law distribution with exponent -1. Let us consider a 2-dimensional lattice and we regard each site as a source of unit income. Initially all the sites on the top line are occupied by different companies, therefore, incomes of all companies are 1. In the second step, a company on the top line is chosen randomly and it occupies a site on the second line, and repeat this process until all sites on the second line are occupied. In the third step a company is chosen randomly from the companies on the second line with the probability weight proportional to the total number of occupied sites on the top and second lines, and let it occupy a site on the third line randomly. Repeat this procedure until all sites on the third line are occupied. The same
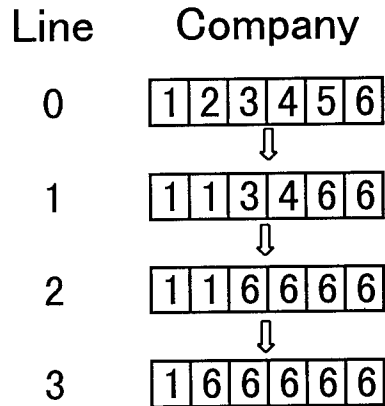
Line Company

0 | 1 | 2 | 3 | 4 | 5 | 6 |
⇓

1 | 1 | 1 | 3 | 4 | 6 | 6 |
⇓

2 | 1 | 1 | 6 | 6 | 6 | 6 |
⇓

3 | 1 | 6 | 6 | 6 | 6 | 6 |

Figure. 9 The company growth model based on competitive growth.
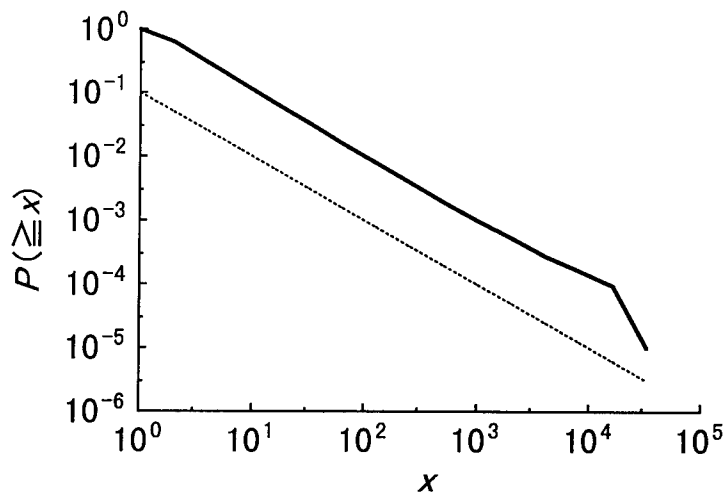

Figure 10 Log-log plot of income distribution by the growth model.

growth process is repeated for many lines (see Fig.10). Note that if a company shares no site on the k-th line, then the company stops growing any more.

In this model each new occupation can be regarded as a unit investment that works. This growth process is based on the intuition that a company's income grows with probability proportional to investment and the amount of investment is assumed to be either 0 or proportional to the whole income. The result of this numerical simulation is shown in Fig.9. We can find a clear power law with exponent −1, which is consistent with Fig.6. Of course this model captures just one aspect of the phenomenon and we need to develop a model that explains both the asset and income statistics.

5　Discussions

We have introduced three topics in this paper; the balance of demand and supply, price fluctuations in open markets and company's macroscopic financial data analysis. These topics may look very different superficially, however, they are deeply interconnected by the key words "fluctuations". Fluctuations had been ignored in economics especially in the discussion of demand and supply, however, as we reviewed in the second chapter the magnitude of fluctuations becomes very large and quick at the balanced point of demand and supply. From the viewpoint of statistical physics the balanced point is the critical point of the underlying phase transition and the fluctuations tend to show fractal characteristics, namely, the averaged amplitude of fluctuations can be infinite in a theoretically ideal situation.

Economic systems tend to be around the critical point in general. Open markets are typical examples. When there are more buyers than sellers, the market price goes up causing decrease in the number of buyers and increase in sellers. Such well-known pulling back mechanism obviously keeps the system around the critical point. The important fact is that this criticality does not stabilize the absolute value of the market price but only stabilizes the statistics of market price fluctuations to follow the critical fluctuations. As the dealers in open markets tend to care about only the relative market price, whether it goes up or down, so the absolute value of the market price is nearly meaningless for the determination of the market price. Therefore, the market prices generally wander almost randomly having statistics characterized by the critical point.

From the viewpoint of fluctuations in demand and supply, company's statistics may also be considered in a similar way. When there appears a new demand in a new field of industry, companies competitively try to supply it. The winner will get a big amount of income and the company will grow. The followers will share the rest of demand and the total supply will increase. When the sum of supply becomes close to the demand, some companies can not gain their incomes and the growth of supply is weakened. As a result demand and supply may nearly balance on average and the whole system may show critical characteristics such as fractal distributions.

The traditional economics theory of demand and supply neglects the effect of fluctuations at all, while the modern financial technology is based on continuous independent Gaussian process. From the standpoint of statistical physics these two cases are extreme limits which correspond to the cases of absolute zero and infinite temperatures, respectively. However, interesting phenomena such as phase transition and pattern formation can never happen in such situations in the case of physical systems. As we have shown in this paper, we can find critical behaviors in real economic data, which considerably resemble physical systems' phase transition phenomenon. It is very likely that there hidden in economic data many more examples which possess fractal properties, because the balance of demand and supply automatically attract any economic system to be around the critical point. Data analysis methods and theoretical formulations developed for phase transition phenomena in physics will become more and more important in economics in the near future.

6　Acknowledgements

References

1. B. B. Mandelbrot, The variation of certain speculative prices, J. of Bussiness (Chicago) **36** (1963) pp.394-419.
2. J. Kertesz and I. Kondor (Eds.), *Econophysics : an emerging science* (Kluwer Academic Publisher, Dordrecht, to appear).
3. R. Mantegna and E. H. Stanley, *Introduction to Econophysics* (Cambridge Univ. Press, London, to appear).
4. H. Takayasu, A.-h. Sato and M. Takayasu, Power law behaviors of dynamic numerical models of stock market prices, (to appear in ref. 2)
5. H. Takayasu and M. Takayasu, Critical fluctuations of demand and supply, *Physica A* **269** (1999) pp.24-29.
6. D. Sornette, D. Stauffer and H. Takayasu, Maket fluctuations II: multiplicative and percolation models, size effects and predictions, (*Procedings book of conference in Rauischholzhausen 1999*, to appear (Eds. A. Bunde) (http://xxx.lanl.gov/abs/condmat /9909439)).
7. H. Takayasu, H. Miura, T. Hirabayashi and K. Hamada, Statistical properties of deterministic threshold elements – the case of market price, *Physica A* **184** (1992) pp.127-134.
8. R. N. Mantegna and H. E. Stanley, Scaling behavior in the dynamics of an economic index, *Nature* **376** (1995) pp.46-49.
9. A.-h. Sato and H. Takayasu, Dynamic numerical models of stock market price: from microscopic determinism to macroscopic randomness, *Physica A* **250** (1998) pp.231-252.
10. H. Takayasu, A.-h. Sato and M. Takayasu, Stable infinite variance fluctuations in randomly amplified Langevin systems, *Phys. Rev. Lett.* **79** (1997) pp.966-969.
11. A.-h. Sato and H. Takayasu, preparing for publication.
12. V. Pareto, *Le Cour d'Economie Politique*, (Macmillan, London, 1896).
13. C. Gini, Indici di concentrazione e di dipendenza, *Biblioteca delli'economista* **20** (1922).
14. E. W. Montrol and M. F. Shlesinger, *J. Stat. Phys.* **32** (1983) 209.
15. M. H. R. Stanley, L.A.N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger and H. E. Stanley, Scaling behavior in the growth of companies, *Nature* **379** (1996) 804.
16. H. Takayasu and K. Okuyama, Country dependence on company size distributions and a numerical model based on competition and cooperation, *Fractals* **6** (1998) pp.67-79.
17. L. A. N. Amaral, S. V. Buldyrev, S. Havlin, M. A. Salinger and H. E. Stanley, Power law scaling in a system of interacting units with complex internal structure, *Phys. Rev. Lett.*, **80** (1998) pp.1385-1388.
18. Y. L. Lee, L. A. N. Amaral, D. Canning, M. Meyer and H. E. Stanley, Universal features in the growth dynamics of complex organizations, *Phys. Rev. Lett.* **81** (1998) pp.3275-3278.
19. K. Okuyama, M. Takayasu and H. Takayasu, Zipf's law in income distribution of companies, *Physica A* **269** (1999) pp.125-131.

FRACTAL APPROACH TO THE REGIONAL SEISMIC EVENT DISCRIMINATION PROBLEM

D.N.BELYASHOV, I.V. EMELYANOVA AND A.V. TICHSHENKO

Institute of Geophysical Research, 191, Abaya Avenue,
Almaty, 480046, Kazakhstan
E-mail: igr@kaskelen.almaty.kz

N. G. MAKARENKO, L. M. KARIMOVA

Institute of Mathematics, 125 Pushkin str.,
Almaty, 480100, Kazakhstan
E-mail: makarenk@math.kz

M. M. NOVAK

School of Mathematics, Kingston University,
Surrey KT1 2EE, England
E-mail: novak@kingston.ac.uk

In the framework of the Comprehensive Test Ban Treaty, development of reliable methods to discriminate between underground nuclear explosions and earthquakes at regional distances (less than 2500km) continues to be very important especially in connection with the last (in May, 1998) nuclear explosions conducted at Indian and Pakistan test sites. Since the lithosphere is a fractal, we suppose the signals, which propagate through the media, inherit its 'self-similar' (scaling) features. We assumed that these features of explosions and earthquakes or their topological reconstructions (embeddings) have to be different. Scaling reflects correlations of more high order then it is possible to estimate by linear discriminating methods and can be used as base of non-linear discrimination. We propose to build a universal geometrical model of a seismic signal using the canon algorithm of F. Takens and to estimate scaling of the model. The scaling features were used as patterns of seismic signals for entering them into an artificial neural network. Records of nuclear explosions and earthquakes from different regions were included into the training set. The net was trained to classify types of seismic events. Results have shown 89% correct classification of the unknown signals. As additional tools for distinguishing between nuclear explosions and earthquakes we propose to use Hurst's method and the cross correlation method. Results of using these methods are demonstrated on examples of some explosions and earthquakes.

1 Introduction

The nuclear explosion discrimination problem continues to be very important especially in connection with the last (in May, 1998) nuclear explosions conducted at Indian and Pakistan test sites. Existing regional methods of seismic event discrimination [1,2,3,4,5] are based on the comparative analysis of spectral characteristics of two main components (P-wave and S-wave) of a seismogram (see Fig. 1). However, these parameters are very sensitive to non-uniformity of the lithosphere and the astenosphere and depend on the location of the event and the path of a signal propagation. Moreover, modern technology of nuclear testing complicates distinguishing between nuclear explosions and earthquakes.

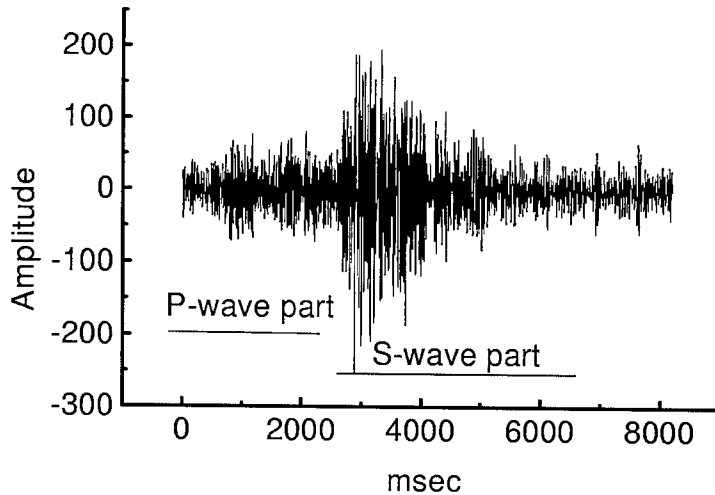It is known that the lithosphere exibites fractal features [6,7] in a wide range of

Figure 1. A seismogram of the Chinese earthquake, 24.07.86

scales - from parts of millimetre of geological grain to tens thousands kilometres of a tectonic plateau and has hierarchy of self-similar blocks. Response of the fractal lithosphere depends on source's geometry (radiation patterns) of a seismic event more than on its yields. Moreover, an earthquake centre is formed in the media during the long time. Features of the media are changed by the moment of the earthquake, which occurs in the preliminary stressed media. An explosion is an artifact for the media and is not connected with any preliminary changes of external parameters of the media. Since dynamic scenarios of explosions and earthquakes are different, we assumed that scaling features of seismic signals or their topological reconstructions (embeddings) have to be different. Scaling reflects correlations of higher order than it is possible to estimate by linear discriminating methods, and can be used as base of non-linear discrimination. We propose to build a universal geometrical model of a seismic signal using the canon Takens' algorithm [8,9] and to use its scaling features as attributes of seismic event discrimination.

The structure of this paper is as follows. In Sec. 2 we study 1D scaling features of seismograms. In Sec. 3 we study correlation dimension of reconstructions of explosions and earthquakes. In Sec. 4 we use the scaling features of embeddings of seismic signals as patterns for a neural network. In Sec. 5 we describe the cross correlation method as a tool for seismic discrimination. Our findings are summarized in the conclusion section.

2 Hurst's exponents of seismograms

We started from the study of self-similar properties of the processes in seismic sources applying Hurst's method [10] to the seismograms of nuclear explosions and earthquakes. Let us remind the classical notion of self-similarity for a random process[11]. $X(t)$ is a self-similar process in $\Re$ if there exists a sequence of positive real numbers c_r such that, for any $r > 0$,

$$X(t) \doteq c_r X(rt), t \in \Re \tag{1}$$

where $\doteq$ denotes equality of all finite-dimensional distribution. This equation is a statement of invariance of $X(t)$ under the group of affine transformations $X \to c_r X, t \to rt, c_r > 0$. Since $c_{a,b} = c_a c_b$ and $c_1 = 1$, c_r must have the form r^H for some $H > 0$ and this formula might be modified as

$$X(t) = r^{-H} X(rt), t \in \Re \tag{2}$$

where H is the Hurst exponent. Traditionally it is estimated by the rescaled range method [10].

We analyzed seismograms of the nuclear explosions conducted in India and Pakistan and Tibetian earthquakes. Fig. 2 represents the Hurst's exponents for different types of seismic events. Analysis of the Hurst's exponents allowed to conclude that graphs of the earthquakes are more regular than graphs of the explosions because H estimated for earthquakes equals 1 on larger range of scales (i.e. $\log T \in [0, 1.5]$) than for explosions (i.e. $\log T \in [0, 1]$).

It is necessary to note that the procedure described above is suitable only for records with low level of noise.

3 Correlation integral as a tool for distinguishing between different seismic sources

The estimation of the correlation dimension [12] of attractors reconstructed directly from experimental time series is often used means of gaining information about the nature of the underlying dynamics. It is proposed that a scalar time series $y(t) = \{y_i\}, i = 1, 2, ..., N$ have been generated by a smooth dynamical system $\mathbf{x}(t + \tau) = \mathbf{f}_\tau(\mathbf{x}(t))$ defined on a manifold M with an attractor $A \in M$ such that $y_i = h(\mathbf{x}(i\tau))$ where $h : M \to \Re$ is Lipschiz function. Then the reconstruction

$$\mathbf{y}(i) = \{y_i, y_{i+1}, ..., y_{i+m-1}\} = $$
$$= \{h(\mathbf{x}(t)), h(\mathbf{x}(t - \tau)), h(\mathbf{x}(t - 2\tau)), ..., h(\mathbf{x}(t - (m-1)\tau))\} \tag{3}$$

defines a *delay-coordinate map* $\Lambda : M \to \Re^m$ and $A \to A_\Re$ is the reconstructed attractor. Takens' theorem [8,9] ensures that if $m > 2D$ where D is the box-counting dimension (or capacity) of A, then the map Λ is embedding, i.e. is one-to-one, and also an immersion with a precision upto assumption about prevalence [13].

For Takens' theorem to be valid, we need to assume that both the dynamics and the observations are autonomous (so that $\mathbf{f}$ and h depend on $\mathbf{x}$ only). Unfortunately,
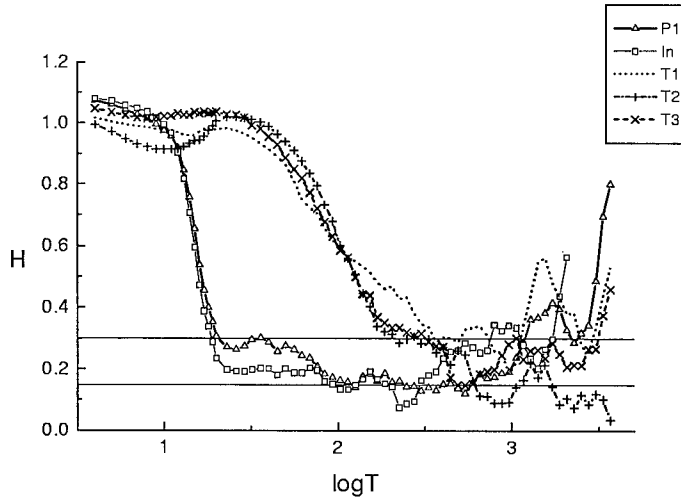
Figure 2. Hurst exponent H curves for different seismic events: Pak1 - nuclear explosion 30.05.98 (Pakistan), ind - nuclear explosion 11.05.98 (India), tib1 - earthquake (Tibet), tib2 - earthquake (Tibet), tib3 - earthquake (Tibet)

these assumptions fail to hold for seismic sources. However, there exists Takens embedding theorem for forced and stochastic systems [14]. We applied the technique of reconstruction described above basing on the last theorem and assuming that the seismograms are typical and contain all the character scales of the dynamics of the seismic sources. It means that in some sense the attractor of the seismic source exists.

We reconstructed some old non-camouflaged nuclear explosions and found out that their embeddings in $\Re^2$ visually differ from ones of earthquakes (see Figures 3-4). Unfortunately, this difference for records of last explosions (Indian and Pakistani explosions) is not visible (see, for example, Fig. 5). Therefore, it is reasonable to use numerous characteristics of these reconstructions. The most popular tool for description of the embedding obtained is the correlation integral [12].

Let $\mathbf{y}(i) \in \Re^m$ be a point on the attractor $A_\Re$. The correlation integral is defined as proportion of pairs of points of no more than distance ε apart. That is,

$$C_m(\varepsilon) = \frac{2}{N(N-1)} \sum_j \sum_k \Theta(\|\mathbf{y}(j) - \mathbf{y}(k)\| - \varepsilon) \tag{4}$$

Here: Θ - Heaviside step function, the symbol $\| * \|$ always denotes 'sup' norm on $\Re^m$.

There is scaling:

$$C_m(\varepsilon) \propto \varepsilon^\nu \tag{5}$$
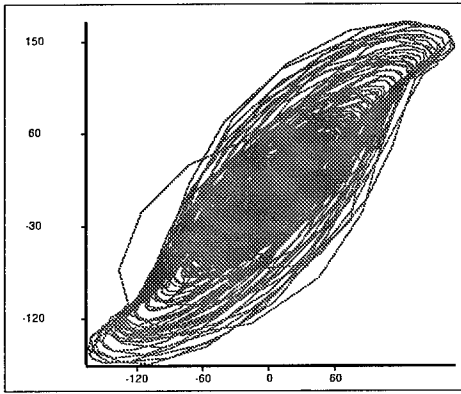
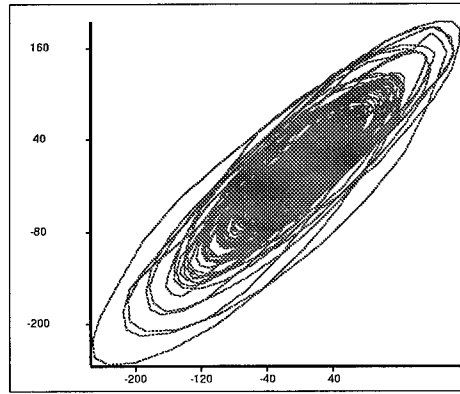Figure 3. Embedding in $\Re^2$, nuclear explosion, STS, 25.04.82.

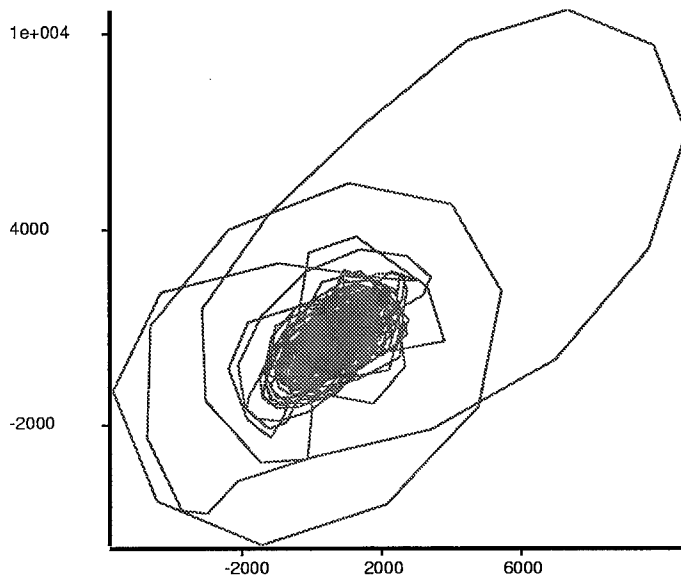Figure 4. Embedding in $\Re^2$, earthquake, China, 24.06.86.

Figure 5. Embedding in $\Re^2$, the Indian nuclear explosion, recorded by NIL station, 30.05.98.

for $\varepsilon \longrightarrow 0$. The slope of the correlation integral is called *correlation dimension*

$$\nu = \lim_{\varepsilon \to 0} \frac{\Delta \log\, C(\varepsilon)}{\Delta \log \varepsilon} \tag{6}$$

The estimated ν typically increases with m and reaches a plateau (in the best case), on which the ν estimate becomes relative constant.

It is expected that because of the spherical symmetry and small sizes of sources

of underground explosions the main part of seismic energy is contained in compressional P-waves having high frequency content in comparison with earthquakes. At the same time earthquakes differ by generating intensive shear S-waves. These differences are traditionally used for choosing diagnostic parameters for seismic event discrimination. Hence, we studied the signals for both types of waves separately.

We processed records of underground nuclear explosions conducted at Semipalatinsk Test Site (STS,Kazakhstan) and Lop Nor (China) and earthquakes in China and Altay (Russia). The seismograms were recorded by Kazakhstani seismic stations BRVK, VOS, ZRN, CHK located in the North-West and TLG - in the South of Kazakhstan. In addition, seimograms of NIL station (Pakistan) were processed.

Our experience showed that correlation integrals for P-wave reconstructions are more informative than ones for S-waves: analyzing correlation integrals for S-waves we didn't find out notable features discriminating different seismic sources. Fig. 7 shows the typical correlation integrals for P-waves of nuclear explosions and earthquakes. It is seen that the correlation dimensions calculated for different types of seismic events are different. As a rule, the correlation dimensions are higher for explosions than for earthquakes.
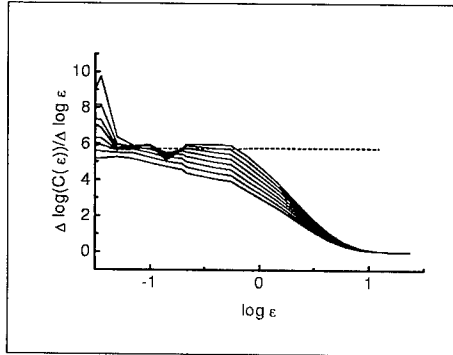
Figure 6. Slopes of correlation integrals ($m = 8 - 16$, $\tau = 2$), P-wave of the nuclear explosion, Semipalatinsk Test Site, 25.04.82.
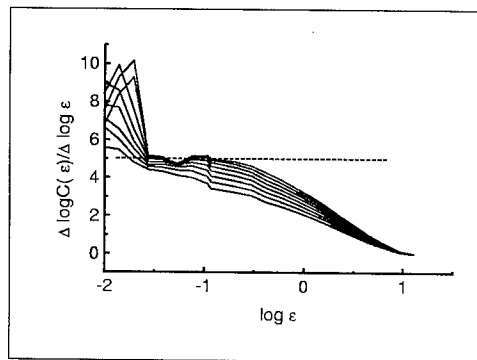
Figure 7. Slopes of correlation integrals ($m = 7 - 14$, $\tau = 2$), P-wave of the Chinese earthquake, 24.07.86.

In general case the slopes exibit complex behavior of ν versus ε Fig. 8. For complex systems it is possible that more than one scaling exists for different ε. The complexity of seismic record slopes might be explained by presence of different kinds folding effects in $\Re^m$ [17]. For the reasons outlined above we used the slope's form instead of the value of the correlation dimension, i.e. the function $\nu = \nu(\log\varepsilon)$.

We used this function, as discriminating attribute of seismic signals for training an artificial neural network to classify seismic sources.
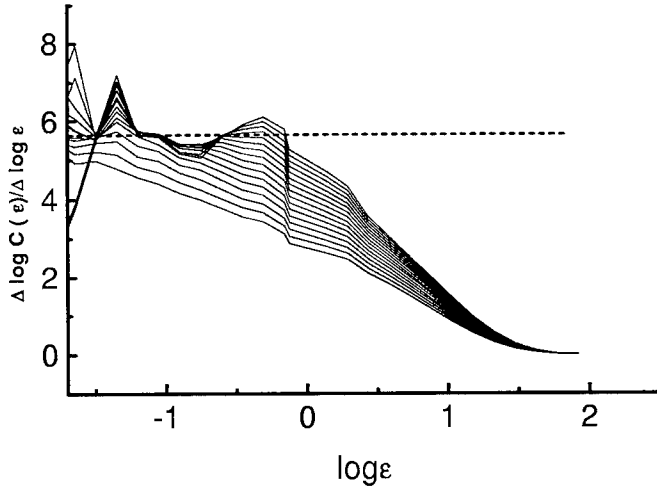
Figure 8. Slopes of correlation integrals ($m = 10 - 25$, $\tau = 2$), P-wave of the Indian nuclear explosion recorded by NIL station 30.05.98.

4 Neural network using

In the last decade artificial neural networks (ANN) have become the most popular tool for pattern recognition tasks [19,20] in general, and in seismic sources distinguishing in particular [3,4,5]. One can consider a neural network as a distributed dynamical system consisting of a set of non-linear processing elements (formal neurons) connected according to a certain architecture. The formal neuron is able to pick up input vectors $\mathbf{x} = \{x_j\}, j = 1, 2, ..., N$, estimate their scalar multiplications with weights $\mathbf{w} = \{w_i\}, i = 1, 2, ..., M$:

$$S = \sum_{i,j} x_j w_i \tag{7}$$

and transform S in accordance with an activation function ϕ into the output vector of the neuron $y = \phi(S)$. Such a neural network might be trained to recognize unknown patterns. Network training process is based on adjusting the weight connections between related values of inputs and targets (desirable outputs) so as to minimize an error function [19]. A set of input-targets values is called a *training* one. There exist two main types of ANN: fully-connected nets and perceptrons [20].

In order to check the algorithm described in the previous section we trained the fully-connected artificial neural network "MultiNeuron" developed by Russian scientists [20] using 60 examples of the correlation integral slopes of seismic records mentioned above as input patterns. The net was tested on 10 seismograms of different underground nuclear explosions and earthquakes which were not included

into the training set. Results of testing showed 89% of the correct classification of the signals.

5 Cross-correlation sums as a tool for distinguishing between different seismic sources

In this section we represent results of our study of non-linear cross correlation sums, which are the generalized correlation integrals and are used here for estimating similarity of two probabilistic measures.

Let $\{y_i\}$, $\{z_i\}$, $i = 1, 2, ..., N$ denote two different observed time series. Internal dynamics of systems in a phase space M which produces these series as typical mapping $M \to \Re$ is unknown. However, it is proposed that there exist chaotic or quasiperiodic attractors of these systems, dimensions of which d_y and d_z are low enough for applying the reconstruction's methods.

Let $A_\Re^{\mathbf{y}}$ and $A_\Re^{\mathbf{z}}$ be their embeddings into $\Re^n$, $n \geq 2(d_y + d_z)$. Let $\mathbf{y}$ and $\mathbf{z}$ be corresponding delay-vectors selected randomly according to two different measures μ and ρ. The cross correlation integrals are defined by[15] :

$$C_{\mu\rho}(\varepsilon) = \int d\mu(\mathbf{y}) \int d\rho(\mathbf{z})\Theta(\varepsilon - \|\mathbf{y} - \mathbf{z}\|) \tag{8}$$

The cross correlation sum is defined similarly by

$$C_{\mathbf{yz}}(\varepsilon) = N^{-2} \sum_{\mathbf{y}} \sum_{\mathbf{z}} \Theta(\varepsilon - \|\mathbf{y} - \mathbf{z}\|) \tag{9}$$

Almost surely, $C_{\mathbf{yz}}(\varepsilon) \to C_{\mu\rho}(\varepsilon)$ for $N \to \infty$, and $C_{\mathbf{yz}}(\varepsilon)$ is an unbiased estimator of $C_{\mu\rho}(\varepsilon)$. For sufficiently small ε and for absolutely continues measures μ and ρ one can show [15]

$$C_{\mu\rho}^2(\varepsilon) \leq C_{\mu\mu}(\varepsilon)C_{\rho\rho}(\varepsilon) \tag{10}$$

In practice, this inequality is used to calculate the cross correlation ratio as a similarity measure

$$r(\epsilon) = \frac{C_{\mathbf{yz}}(\varepsilon)}{\sqrt{C_{\mathbf{yy}}(\varepsilon)C_{\mathbf{zz}}(\varepsilon)}} \tag{11}$$

We used it in the form [16]

$$K_{\mathbf{yz}}^m(\varepsilon) = \sqrt{\frac{\sum_{i \neq j} \|\mathbf{z}(i) - \mathbf{z}(j)\|^2 \Theta(\varepsilon - \|\mathbf{y}(i) - \mathbf{y}(j)\|)}{\sum_{i \neq j} \Theta(\varepsilon - \|\mathbf{y}(i) - \mathbf{y}(j)\|)}} \tag{12}$$

If $\{y_i\}$ and $\{z_i\}$ are related by identical dynamical scenarios than one can expect that $\|\mathbf{y}(i) - \mathbf{y}(j)\| < \varepsilon \Rightarrow \|\mathbf{z}(i) - \mathbf{z}(j)\| \approx \varepsilon$. If no, $\|\mathbf{z}(i) - \mathbf{z}(j)\| \approx \Delta z$, where Δz is the average distance between points of the attractor $A_\Re^{\mathbf{z}}$ and $K_{\mathbf{yz}}^m$ does not depend on ε.

Fig. 9 represents the cross corelations of various seismic events. The cross correlation between the Tibetian earthquake (09.01.90) and the Indian nuclear explosion (11.05.98) is absent: $K_{\mathbf{yz}}^m(\varepsilon)$ curve is practically in parallel with the horizontal axis.
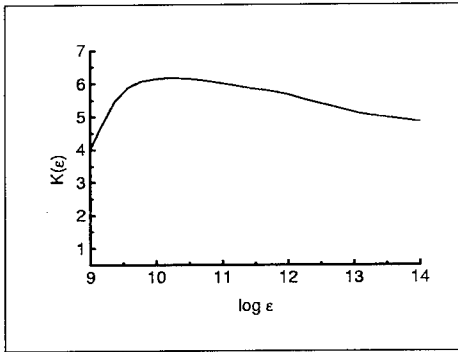
Figure 9. Cross correlation between reconstructions of the Tibetan earthquake and the Indian nuclear explosion.
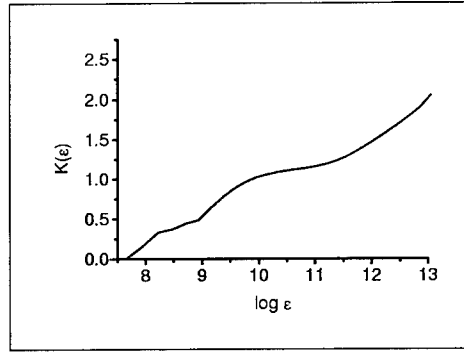
Figure 10. Cross correlation between reconstructions of the Indian and the Pakistani nuclear explosions.

The Figure 10 demonstrates strong cross correlation between the Indian explosion (11.05.98) and the Pakistani nuclear explosion (28.05.98). In this case we can conclude that dynamical scenarios of two systems observed are identical ones.

6   Conclusion

We have introduced non-linear criteria for distinguishing between underground nuclear explosions and earthquakes recorded at regional distances using the fractal approach.

We noted that behavior of Hurt's exponents varies for graphs of different seismograms in dependence on scale: the regularity range for the earthquakes is larger than for the explosions.

We found out that embeddings of non-camouflaged explosions into $\Re^2$ visually differ from ones of earthquakes. The slopes of correlation integrals demonstrate complexity of scaling. The notable plateau is absent for both types of seismic sources but the curves of the slopes are different for explosions and earthquakes.

The correlation dimension curves of 60 seismograms of different seismic events were entered into the neuroimitator "MultiNeuron", which was used as a classificator of underground nuclear explosions and earthquakes. The testing results obtained have shown 89% correct classification of 10 seismic events, which were not included into the training set.

Sometimes, it is possible to identify the seismic source estimating its cross correlation with a test example. Analysis of relationships of attractors in the phase space allows obtaining additional information for the discrimination task.

Therefore, our experience has shown that the non-linear methods could be successfully applied to regional seismic event discrimination.

268

References

1. P.W.Pomeroy, W.J.Best, T.V.McEvilly, Bull.Seism Soc.Am. **72**, 89 (1982)
2. S.R.Taylor, M.V.Denny, E.S.Vergino,and R.E.Glaser, Bull.Sesm.Soc.Am. **79**, 1142 (1989)
3. F.Dowla, S.Taylor and W.Anderson, Bull.Sesm.Soc.Am. **80**, 1346 (1990)
4. Y.Shimshoni, N.Intrator in Proc. of ICONIP'96 Springer, Hong Kong1861996
5. P.Dysart and J.Pulli, Bull.Sesm.Soc.Am. **80**, 1910 (1990)
6. M.A.Sadovskii, Doklady, Academy of Sciences of the USSR **247**, 829 (1979)
7. M.A.Sadovskii,V.F. Pisarenko, V.N. Rodionov, Herald of the Academy of Sciences of the USSR **N1**, 82 (1983)
8. F. Takens, in Lecture Notes in Math. **898**, 336 (1981)
9. L.Noak, Inter.J. of Bifurcation and Chaos **1**, 867 (1991)
10. J.Feder *Fractals* (Plenum Press, New-York, 1988)
11. D. Veneziano, Fractals **7**, 59 (1999).
12. M. Ding, C. Grebodi, E. Ott E., T. Sauer , J.A. York, Physica D **69**, 404 (1993)
13. T. Sauer, J.A. Yorke, M.Casdali, Journal of Statistical Physics **65**, 579 (1991).
14. J. Stark, D.S. Broomhead, M.E. Davies and J. Huke, Nonlinear Analysis, Theory, Methods and Applications **30**, 5303 (1997)
15. P. Schneider, P. Grassberger, Studying Attractor Symmetries by Means of Cross Correlation Sums. Available on http://babbage.sissa.it/chao-dyn/9604020
16. G. Lasiene and K. Pyragas, Physica D **52**, 332 (1991).
17. A. Potapov, J.Kurths, Physica D **120**, 369 (1998).
18. D.N.Belyashov, I.V.Emelyanova, A.V.Tichshenko, N.G.Makarenko and L.M.Karimova in Neuroinformatics-99 Moscow Institute of Enginering and Physics, Moscow, Russia, 3,31,1999, in Russian.
19. C.M. Bishop *Neural Network in pattern recognition* ed. (Oxford University Press, Bristol, 1996)
20. A.N.Gorban' and D.A. Rossiev *Neural Network on a personal computer* (Nauka, Novosibirsk, 1996), in Russian.

FRACTAL AND TOPOLOGICAL COMPLEXITY OF RADIOACTIVE CONTAMINATION

N. G. MAKARENKO ,L. M. KARIMOVA

Institute of Mathematics, 125 Pushkin str.,
480100, Almaty, Kazakhstan
E-mail: makarenk@math.kz

A. G. TEREKHOV

Institute of Space Research, 26 Shevchenko str.,
480100, Almaty, Kazakhstan
E-mail: remote@astel.kz

M. M. NOVAK

School of Physics, Kingston University,
Surrey KT1 2EE, England
E-mail: novak@kingston.ac.uk

There is verified the hypothesis about multifractal nature of radioactive contamination due to nuclear explosions on Semipalatinsk test site (STS) in Kazakhstan. The fields of terrestrial contamination have extreme high variability caused by a number of nature processes. There are used the methods of multifractal formalism and computation topology. The results of analysis support the existence of multifractal scaling, which is different for natural and man-made isotopes. It means that there are "hot spots" of man-made Cs isotope contamination dangerous for human health, which can't be detected by the traditional in Kazakhstan technics of measurements from air. Moreover, the finding of self-similarity could be the base of new methods of diagnostics of large territories.

1 Introduction

There were some 470 nuclear explosions on STS in Kazakhstan, of which 90 explosions were in the air, 25 on the ground and 355 underground. As a result a lot of territories were contaminated and it has become a matter of serious for people with regard to its consequences for the lives and health [1]. We analyze the resulting contaminated areas by examining the distribution of radionuclides. The analysis of the problem is not trivial and is influenced by a number of issues, such as wide range of contamination, extending from meters to hundreds of kilometers; presence of a mixture of radioactive isotopes, due to numerous nuclear explosions of various composition and power; highly variable rate of migration of radionuclides in soils; the influence of climatic conditions on the localisation of contamination and its transport [2]. So, the radioactive contamination of the ground is a result from nonlinear interplay of geophysical fields which intervene over a large range of scale. Consequently, fields of contamination observed have an extreme spatial variability, frequently cited "hot spots" or "leopard's skin" [3].

Complex structure of the field is the main difficulties of it's diagnostics. The absence of the field smoothness restricts a choice of methods of measurements and data processing. The maps of contamination are designed using interpolation of

measurements and, consequently, depend on a scale of averaging. As a result, sparse anomalies can be missed.

To detect all fluctuations of the contamination field measurements must cover a territory by a network of overlap neighborhoods[4]. But it is impossible for huge areas of former nuclear weapon testing site, spanning an area of around $18500 km^2$ and regions bordering this test site which we deal with. Nowadays, only 1% of contaminated territory has been investigated by accepted in Kazakhstan methods of measurements from air and it is definitely not enough to describe the radioactive contamination and to implement effective decontamination methods. The scale-symmetry assumption is the simplest and also only one acceptable in the absence of knowledge in specific mechanisms producing these fields. The idea of scaling invariance can be useful in practice and could become a base for the analysis and diagnostics of the areas.

The present study deals with the area-analysis of the distribution of radionuclides in Kazakhstan. The data is obtained primarily from former STS and, secondly, from regions bordering this test site. Experimental data consist of radionuclide measurements made from the air, on a grid of parallel lines covering a part of territory. Two regions, Karaganda and Semipalatinsk, and Irtysh test site were explored.

It is possible to analyze these measurements by several complementary methods, among them the fractal [5], multifractal [3,6] and morphology methods [7]. Here, we present multifractal method [8,9] and method of computational topology [10].

The structure of this paper is as follows. In Sec. 2 we recall some definitions of multifractal analysis and apply this technique to the data from flight paths. In Sec. 3 we describe the method of computational topology and present the results of applying this method and multifractal one to data from the Irtysh test site. The summary is found in the concluding section.

2 Multifractal analysis of radioactive contamination

We used an approach in which the contamination data along a path are considered as multifractal random measures. Let us remind [8,9,11], that multifractal spectrum of singularities α of Borel finite measure μ on a *compact* set X is a function $f(\alpha)$ defined by pair (g, G). Here, $g : X \to [-\infty, \infty]$ is a function, which determines the level sets:

$$K_\alpha^g = \{x \in X : g(x) = \alpha\} \tag{1}$$

and produces a multifractal decomposition X:

$$X = \bigcup_{-\infty \le \alpha \le \infty} K_\alpha^g. \tag{2}$$

Let G be a real function which is defined on $Z_i \subset X$ such as $G(Z_1) \le G(Z_2)$ if $Z_1 \subset Z_2$. Then multifractal spectrum is $f(\alpha) = G(K_\alpha^g)$. Let g be determined as pointwise dimension d_μ of measure $\mu(x)$ at all points $x \in X$ for which the limit

$$g \equiv d_\mu(x) = \lim_{r \to 0} \log \mu(B(x,r))/\log r, \tag{3}$$

exists, where $\mu(B(x,r))$ is a "mass" of measure in the ball of radius r centered at x. Since, we have chosen $g = d_\mu$, we can drop the g subscript from further references to K_α^g. Then $K_\alpha = \{x : d_\mu(x) = \alpha\}$, where exponent α is local density of μ. The singular distribution μ can then be characterized by giving the "size" of sets K_α by their Hausdorff dimension, i.e. $f(\alpha) = G(K_\alpha) = dim_H(K_\alpha)$. If μ is self-similar in some sense, $f(\alpha)$ is a well behaved concave function of α [9]. To estimate $f(\alpha)$ for path data we applied the method of the partition sum [11].

Let $X = \Im = [0,1]$ and k_n be an increasing sequence of positive integers. Define:

$$I_{i,n} = \left[\frac{i}{k_n}, \frac{i+1}{k_n}\right]. \tag{4}$$

The component of multifractal decomposition $\Im$ is

$$K_a = \{x \in \Im \mid \lim_{n \to \infty} \frac{\log \mu(I_{i,n}(x))}{\log k_n} = \alpha\}. \tag{5}$$

It is more easy to analyze "coarse grained" Hölder exponents [12] $\alpha = \alpha_n$ of μ on $I_{i,n}$:

$$\alpha_n(I_{i,n}) = \frac{\log \mu(I_{i,n})}{\log k_n}. \tag{6}$$

Their variations are described by Legendre spectrum [13]:

$$f(\alpha) = f_L(\alpha) = \inf_q(\alpha q - \tau(q)), \tag{7}$$

where

$$\tau(q) = \lim_{n \to \infty} \frac{1}{\log k_n} \log \sum_i [\mu(I_{i,n})]^q \tag{8}$$

and summation runs through i such that $\mu(I_{i,n}) \neq 0$.

We analyze the radionuclide contamination data measured from the air for Karaganda and Semipalatinsk regions. The height of all flights was about 50 m and the velocity of the plane was about 125 km/h. Spectrometrical measurements were usually made at time intervals of 1 sec. The technique used is based on measurements of γ-quanta flow density of ^{214}Bi (1.12 and 1.76 Mev) to determine the contamination by U, ^{208}Tl (2.62 Mev) to determine Th contamination, ^{40}K (1.46 Mev) to determine K and ^{137}Cs (0.66 Mev) for Cs contamination. Total γ-activity is measured in the range of 0.25-3.0 Mev. The spectrometer was equipped by NaI element. The measurements have been done at different scales, e.g., in Karaganda and Semipalatinsk region the scale is 1:1,000,000, so the distance between paths is 10 kms. There are up to 11000 data points along the paths, with each reading separated by 50 meters. A sample of measured data can be found in Fig. 1.

For calculation of multifractal spectrum it is necessary to transform contamination values Y into probabilistic measures. As a rule, empirical probability density function $p(Y)$ behaves itself as a curve (1) shown in Fig. 2, where function tails fall down. The path data Y were transformed to new variables Z which have rectangular random probability density function (the curve 2 in Fig. 2) by solution of
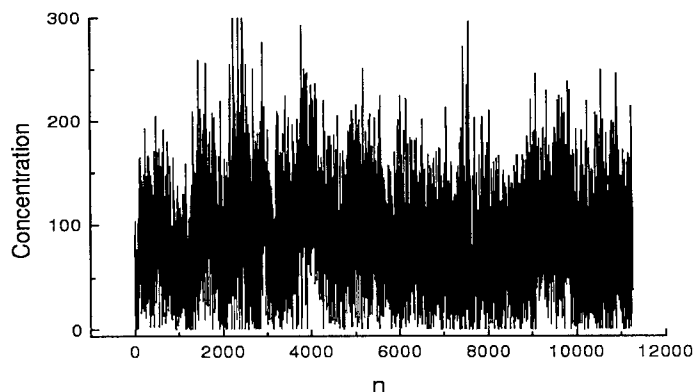
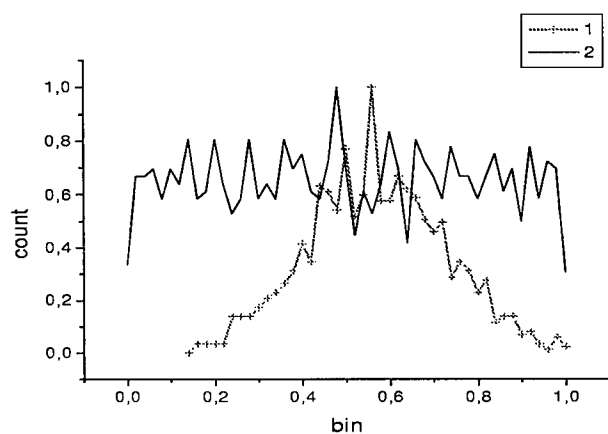Figure 1. Data array of Cs, Semipalatinsk region path

Figure 2. The histograms of distribution contamination data (1) and probabilistic measures (2).

equation $P(Y) = P(Z)$, where P is a probability function. In the case of rectangular random distribution $Z = P(Y)$. This transformation is equivalent to coordinate change which saves a dimension spectrum $f(\alpha)$ [14].

The range of q value was chosen as $[-6, 4]$. The Fig. 3 displays the set of multifractal spectra of radionuclide contamination of paths. The scaling of all isotopes obeys to multifractal law. The behaviors of $f(\alpha)$ curves are different for various isotopes. In Fig. 4 it is shown the set of $f(\alpha)$ spectra of man-made isotope Cs data calculated for different paths. It can be seen that scaling properties differ from path to path, but this scaling variability is not so high as distinction of $f(\alpha)$
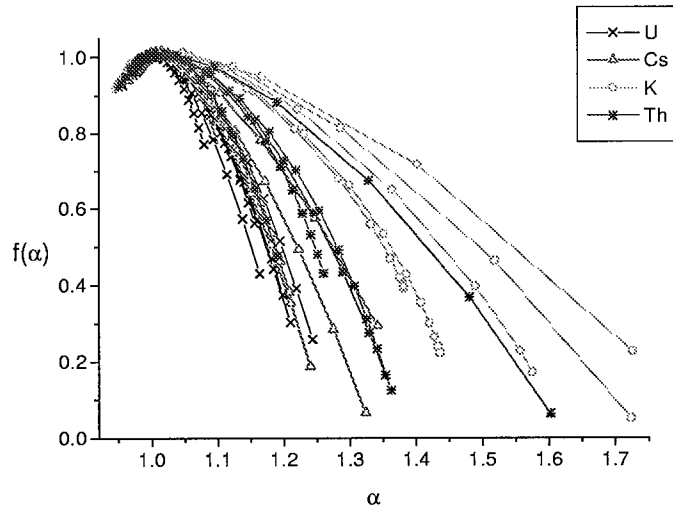
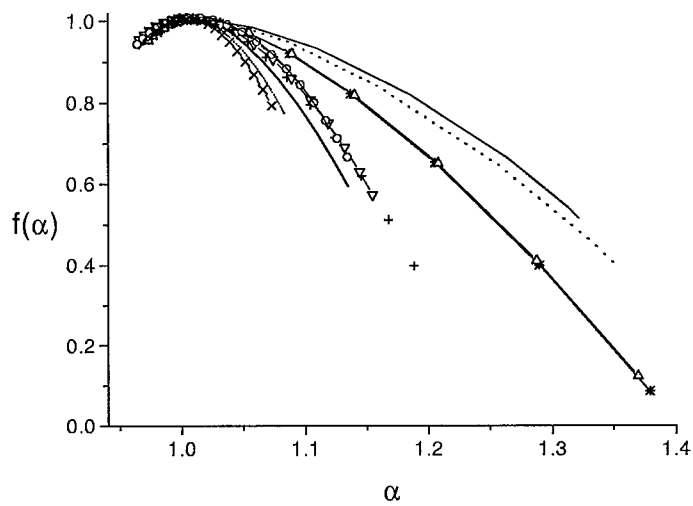Figure 3. The set of $f(\alpha)$ spectra of Karaganda path contamination.

Figure 4. The set of $f(\alpha)$ spectra of Cs data for some paths.

spectra for different isotopes (cf. Fig. 3).

The results obtained by multifractal tool are related to separate paths, where

a distance between adjacent paths is $\sim 10km$ and effective strip of equipment capture is $\sim 300m$. For the case of measurements paving the territory we can use the method of computational topology described in the next section, which complements multifractal method.

3 Computational topology of the radionuclide fields: disconnectedness index.

The multifractal formalism can be applied to area contaminations, where a support of measure $\mu(x,y)$ is a compact set $X = [0,1] \times [0,1] \in \Re^2$. In this case [13] $I_{i,n} \to I_{i,j,n}$, where

$$I_{i,j,n} = \left[\frac{i}{k_n}, \frac{i+1}{k_n}\right] \times \left[\frac{i}{k_n}, \frac{i+1}{k_n}\right]. \tag{9}$$

Unfortunately, contamination data do not allow to use square paving, as the distance between the data points on adjacent paths is always larger than the distance between the neighbouring points along a single path. It could be used the tesselation of the surface by rectangles $\{c_i\}$ [15]. In this case probability measure on c_i is averaged over rectangle, such as, $\mu(x,y) : x,y \in c_i$ presents itself step-like function, which has a non-zero constant value at each rectangle. The components of "coarse-graining" multifractal decomposition K_α are found by selecting of rectangles c_i of Hölder exponent $\alpha \pm \Delta\alpha$, where $\Delta\alpha$ is discrete step of the measure. The coarse Hölder exponent is

$$\alpha(c_i) = \frac{\log \mu(c_i)}{\log |c|}, \tag{10}$$

where $|c|$ is a size of c_i. "Coarse-graining" version $f(\alpha)$ is

$$f_G(\alpha) = \lim_{\Delta\alpha \to 0} \lim_{\delta \to |c|} \frac{\log N_\delta(\alpha, \Delta\alpha)}{\log(1/\delta)}, \tag{11}$$

where $N_\delta(\alpha, \Delta\alpha)$ denotes the number of boxes of δ size with $\alpha(c_i) \simeq \alpha$. This function can be explained in statistical terms [12] as probability p_δ of finding $\alpha(c_i) \simeq \alpha$ and behaves roughly like

$$N_\delta(\alpha)/N_\delta = p_\delta\left[\alpha(c_i) \simeq \alpha\right] \simeq \delta^{D - f_G(\alpha)}. \tag{12}$$

Here, N_δ is the total number of non-empty δ-boxes which contain measure, D denotes the box dimension of the support of μ.

For estimation of $f_G(\alpha)$ spectrum it could be used functional box-counting [15], however the covering by rectangle neighborhoods makes it impossible to re-scale rectangle edges uniformly and introduces additional local and global dimensions which have no clear meaning [16]. Consequently, we use computational topology for calculation of the box dimension $f_G(\alpha)$.

The component K_α of multifractal decomposition is a subset $\{c_i\}$ having the same $\alpha \pm \Delta\alpha$. The "coarse-graining" geometry of K_α can be described by the rate of growth a number of $c_i \in K_\alpha$ with improving resolution [10] and called *disconnectedness index* γ.
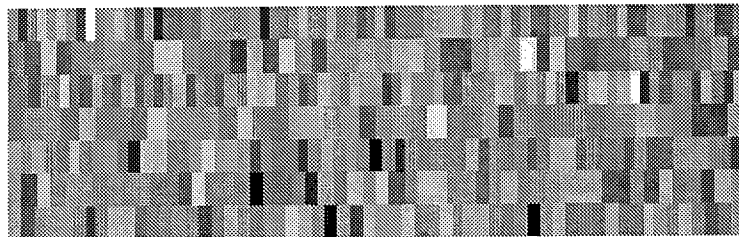
Figure 5. The paving map of U contamination of Irtysh test site.

A topological space X is connected if and only if it cannot be decomposed into the union of two non-empty, disjoint, closed sets. If such a decomposition exists then X is said to be disconnected– that is, if there are two closed sets U and V such that $U \bigcap V = 0$ and $U \bigcup V = X$. There was reformulated the notion of connectedness in order to make it possible to implement a test for this property on a computer. The basic idea is to look at the set with a finite resolution ϵ and see how connectedness changes as we let $\epsilon \rightarrow 0$. It is said that a subset X of metric space is $\varepsilon - disconnected$ if it can be decomposed into sets that separated by a distance of at least ϵ.

Let $N_\varepsilon(\alpha)$ be the number ε-components by given a resolution ε and intensity of measure equals $\alpha \pm \Delta\alpha$. The disconnectedness index γ is the following limit:

$$\gamma = \liminf_{\varepsilon \to 0} \frac{logN_\varepsilon(\alpha)}{log(1/\varepsilon)}. \tag{13}$$

The index equals to Hausdorff dimension for the simplest sets, for example for the middle-thirds Cantor set. However, more generally the disconnectedness index distinguishes between sets with the same box-counting dimension.

3.1 Scaling properties of Irtysh area contamination.

There have been applied both multifractal formalism and computational topology to Th, K, U, Cs isotope contamination of Irtysh area which is the part of STS. Irtysh area was chosen for calibration of the devices for spectrometrical measurements from air and was investigated more thoroughly. The distance between paths was 50 m, effective strip of equipment capture was $\sim 50m$ and there were 1197 measurements on the ground.

To calculate γ index a range of values for each isotope have been divided into 255 levels. The contamination values $\nu(c_i)$ are expressed in the share of the whole range which was reduced to 1. In Fig. 5 a part of the digital image of contamination used for γ computation are shown.

The curves of disconnectedness index (Fig. 7) of natural isotopes have plateaux with $\gamma = 2$ at $\nu \in [0.3 \div 0.5]$ for Th, $\nu \in [0.3 \div 0.45]$ for K and $\nu \in [0.3 \div 0.35]$ for U and the γ values decrease as ν increases. The behavior of man-made Cs isotope is different. The index of disconnectedness γ is equal to 1.3 at a small ν and slowly decreases demonstrating two plateaux: $\gamma = 1$ at $\nu \in [0.45 \div 0.7]$ and plateau $\gamma \approx 0.6$ at $\nu > 0.8$ (438 milliCurie/km^2). The latter means the possibility of existence of
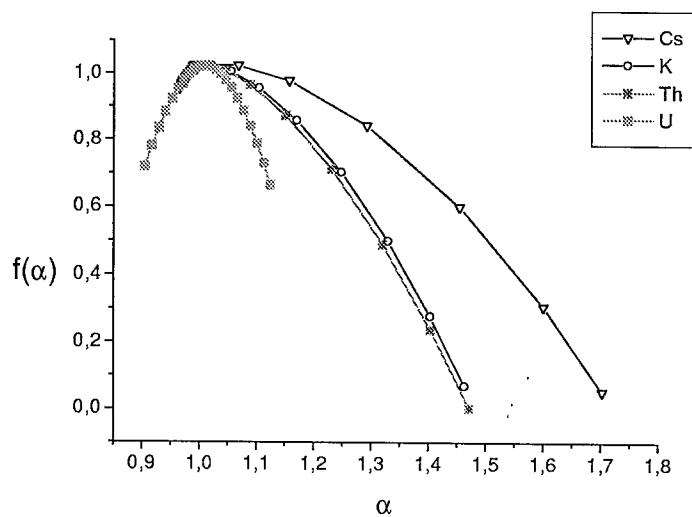
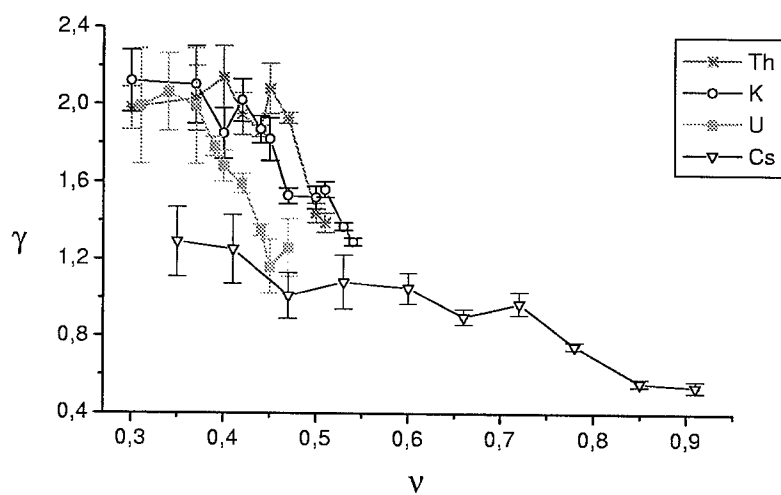Figure 6. The $f(\alpha)$ spectra of radionuclide data for Irtysh test site.

Figure 7. The dependence of disconnectedness index versus cross-section level for Th,K,U and Cs isotopes.

"hot spots" of Cs contamination which lie in the area of highest concentrations ($> 438 milliCurie/km^2$) and small fractal dimensions $\gamma < 0.6$. In general, $\gamma(\nu)$ can be interpreted as "pre-multifractal" spectrum $f_G(\alpha)$. In this case, the probability of existence of "hot spots" decreases with spot size δ as $\propto \delta^{(2-\gamma)}$ (cf. Eq. 12). It is necessary to notice the similar behavior of Cs cumulative soil deposition data in Chechslovakia after Chernobyl accident [6].

Statistical properties of multifractal fields are a function both of a scale and a dimension of a support of measurements. It is known, that measurement net with empirical box-dimension D_{net} cannot detect sparsely distributed phenomena with box-dimension $D_a < 2 - D_{net}$ [17]. This conclusion follows from the common ideas about *tranversality* [18]. It is necessary remember, that two sets M and N of dimension D_M and D_N intersect *transversally* in $\Re^m$ if

$$codim(M \cap N) = codim D_M + codim D_N. \qquad (14)$$

So, the net with dimension D_{net} can detect a set of dimension D_a only if $D_{net} + D_a > m$. As a rule, box-dimension of the net of spectrometrical measurements from air is $D_{net} \ll 2$, consequently, the number of anomalies on contaminated territories can be missed.

In addition to computation topology method there was used multifractal formalism for estimation of $f(\alpha)$ spectra of Irtysh test site data. There was constructed one dimension array by linking all paths in one. Adjusted paths were "glued" together by taking into account their orientation and adding the beginning of the following path to the end of previous one. Such arrays were formed for all isotopes and multifractal spectra were computed. The results of this analysis are shown in Fig. 6. The comparison of results of both methods applied, to Irtysh data, points at good concurrence of isotope discrimination on their scaling. The qualitative behaviour $f(\alpha)$ and $\gamma(\nu)$ (Fig. 7) curves are the same. Man-made isotope Cs has demonstrated the most large Hölder exponent $\alpha \sim 1.7$ in multifractal spectra. It means that there are hot spots of Cs contamination which form a sparse set having dimension $f(\alpha) < 0.2$.

4 Conclusion

All the results of the multifractal analysis that we obtained support the existence of multifractal nature in the terrestrial radionuclides contamination in the two investigated regions and Irtysh area. The analysis made it possible to distinguish radionuclide isotopes K, Th, U, Cs, specially for measurements made on Irtysh test site.

Using the method of computation topology for the case of measurements paving the territory, a pre-multifractal spectrum of dimensions (indexes of disconnectedness) for Irtysh area was calculated. It was found that the behavior of disconnectedness index of natural isotopes differs strongly from Cs one. Contamination of natural radionuclides exhibits $\gamma \approx 2$ for low values and dimension decreases quickly with increasing of level. Spectrum of dimensions for Cs points out the existence of "hot spots" of contamination.

Taking into account the multifractal properties of man-made Cs found for Irtysh

area, it can be pointed out the existence of dangerous for human health anomalies on all contaminated territories. These "hot spots" cannot be detected by using spectrometrical measurements from air accepted in Kazakhsatan. The results obtained can be the background of new methods of diagnostics of large territories.

Acknowledgments

References

1. Resolution of General Assembly of United Nations **A/Res/53/1D**, 7 (1998)
2. Yu.A. Izrael, Herald of the Russian Academy of Sciences **68**, 898 (1998) (in Russian)
3. G. Salvadori, S.P. Ratti and G. Belli, Health Physics **72**, 60 (1997).
4. M. Reginatto, P. Shebell and K.M. Miller, in *EML-590* ed. Environmental Measurements Laboratory, Department of Energy, U.S., October 1997
5. N. Makarenko, L. Karimova, and M.M. Novak, Fractals **6**, 359 (1998).
6. G. Salvadori, S.P. Ratti and G. Belli, Chemosphere **33**, 2359 (1996).
7. N. Makarenko, L. Karimova, and M.M. Novak, in *Proceedings of Fractals in Engineering* ed. (INRIA, 1999)
8. L. Barreira, Y. Pesin and J. Schmeling, Chaos **7**, 27 (1997)
9. K.J. Falconer, Jour. of Theoret. Prob **7**, 681 (1994)
10. V. Robins, J.D. Meiss and E. Bradley, Nonlinearity **11**, 913 (1998).
11. R.H. Riedi and J. Lévy Véhel, IEEE Transactions of Networking **3129**, 1 (1997)
12. R.H. Riedi and I.Scheuring, Fractals **5**, 153 (1997)
13. J. Lévy Véhel and J.-P. Berroir, Fractals in the Natural and Applied Sciences, IFTP **A-41**, 261 (1994)
14. M. Holschneider, Fractal Dimensions: A New Definition, M-CPT **2140**, 1 (1988)
15. S. Lovejoy and D. Schertzer and A.A Tsonis, Science **235**, 1036 (1987)
16. B.B. Mandelbrot, in *Fractals in Physics* ed. L. Pietronero and E. Tossati, North-Holland, Amsterdam, 1986
17. S. Lovejoy and D. Schertzer and P. Ladoy, Nature **319**, 43 (1986)
18. D. R. J. Chillingworth, in *Differential topology with a view to applications* ed. (Pitman Publishing London, San Francisco, Melbourne, 1976)

PATTERN SELECTION: NONSINGULAR SAFFMAN-TAYLOR FINGER AND ITS DYNAMIC EVOLUTION WITH ZERO SURFACE TENSION

HUIDAN YU

*Department of Physics, Zhejiang Normal University, Jinhua321004, Zhejiang, China
and Laboratory for Nonlinear Mechanics of Continuous Media, Institute of Mechanics,
Chinese Academy of Sciences, Beijing, China,
and Department of Physics, Peking University, Beijing 100871, China*

KAIHUA ZHAO

Department of Physics, Peking University, Beijing 100871, China

By a slight modification of Saffman-Taylor's viscous finger, we remove its singu-
larity and present a finite-length nonsingular Saffman-Taylor finger. Its dynamic
evolution solutions, both one-step theory prediction and numerical simulation in
the long-time limit with zero surface tension, give an alternative proof for the an-
alytical selection demonstration of the Saffman-Taylor finger width in the absence
of surface tension. As a simple example of pattern selection, which has real exper-
imental background, this work not only contradicts the generally accepted belief
that surface tension is indispensable for the selection of the $\frac{1}{2}$-width finger but also
provides more models to compute the evolution, competition and ramification of
multiple fingers numerically in straight channel as well as circular disc geometry.

1 Introduction

The Saffman-Taylor problem [1] has played a central role in the study of viscous
fingering in a Hele-Shaw cell [2], which was modeled by two-dimensional potential
flow at the interface between two fluids. Saffman and Taylor found analytically a
continuous family of steady-state solutions, which shows fingers of different width
could exist in the absence of surface tension, but their experiments with negligible
surface-tension effect in 1958 and numerical calculations made by McLean, Saffman
and Vanden-Broeck in the presence of surface-tension effect in 1981 and 1983 showed
no hint of the continuous family [3,4]. On the contrary, the finger with relative width
$\lambda = 0.5$ is always singled out at the zero surface tension limit. This problem
(which is so called Saffman-Taylor) has been of much subsequent interest because
it is universal, i.e., the same selection phenomenon is common for displacement of
various viscous liquids by less viscous ones for immiscible incompressible liquids.
Much work [5] was done toward solving this Saffman-Taylor puzzle. However, it
has been widely accepted that the inclusion of surface tension is the only way
to select the most stable finger width since Saffman-Taylor proposed that surface
tension between the two fluids would solve the selection problem [1]. Although several
works [6] confirmed numerical evidence of the discrete spectrum of λ, decreasing to $\frac{1}{2}$,
in the limit of low surface tension, exact solutions are rare with zero surface tension.
Recently, Mineev-Weinstein [7] analytically solved the finger selection problem in the
absence of surface tension. He showed an exact result that a generic interface in
a Hele-shaw cell evolves to nonlinearly stable single uniformly advancing finger
occupying one-half of the channel width, which contradicts the generally accepted
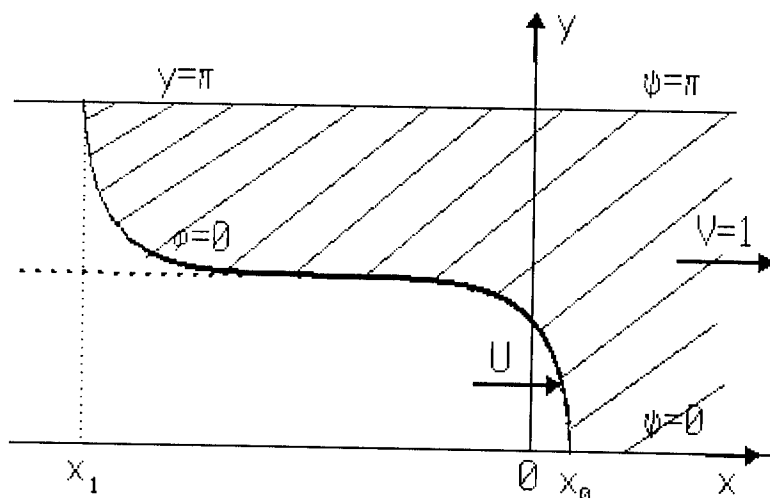
Figure 1.

belief that surface tension is indispensable for the selection of the $\frac{1}{2}$-width finger.

We find that the problem of the finger width selection stems from the singularity of Saffman-Taylor finger, which is half infinite in length. The singularity makes the numerical calculation impossible and is not in agreement with the experimental results. We introduce a positive variable $\epsilon(t)$ to remove the singularity. By the conformal mapping method presented by Bensimon [8], we predict that there exists a unique steady width $\lambda = 0.5$ of the viscous finger to which the fingers of other width approach in the absence of surface tension after a one-step analysis. Then we verify this prediction by numerical computation and show its long time evolution.

2 Nonsingular Saffman-Taylor Finger and One-step Dynamic Evolution Theory

Following Saffman and Taylor we consider the analysis of a long air bubble or finger moving through a channel in the Hele-Shaw cell bounded by two straight parallel walls. We take the y-axis (Fig. 1) perpendicular to the walls and the x-axis along the central axis of symmetry of the channel. The separation of the walls is normalized to 2π and the velocity of the fluid at infinity in front of the finger is normalized to unity. Let $z = x + iy$ and $\Phi = \phi + i\psi$, $\phi(z)$ and $\psi(z)$ being the velocity potential and stream function, respectively.

In case of the absence of surface tension, the interface is described by $\phi = 0$, which can be conformally mapped to the circumference of a unit disk in the complex

plane $\zeta = e^{-\Phi}$. By these notations the Saffman-Taylor solution [1] can be written as

$$z = 2(1 - \lambda)ln\frac{1}{2}(e^{-i\psi} + 1) + i\psi, \tag{1}$$

where the single parameter λ depicts the width of the finger. Saffman-Taylor finger is half infinite in length and the solution is steady in the frame co-moving with the finger tip which has a speed $\frac{1}{\lambda}$ relative to the laboratory frame.

In order to remove the singularity at $\psi = \pm\pi$ in Eq. (1), we choose the solution in the following form in the laboratory frame:

$$z = 2(1 - \lambda)ln\frac{1}{2}(\zeta + 1 + \epsilon) - ln\zeta + \frac{t}{\lambda} + \widehat{z}(\zeta, t), \tag{2}$$

where $\epsilon = \epsilon(t)(=\epsilon_0$ at t=0$)$ is a small positive real number varying with time t, $\widehat{z}(\zeta, t)(=0$ at t=0$)$ describes the possible additional time-dependence of the finger. The functional form of $\epsilon(t)$ and $\widehat{z}(\zeta, t)$ are to be determined by the evolution equation of interface. In Eq. (2) the singularity of Saffman-Taylor solution at $\zeta = -1$ on the unit circle is moved to outside of the unit disk and therefore the length of the finger becomes finite (Fig.1).

The corresponding equation of the boundary of the nonsingular Saffman-Taylor finger is

$$z = 2(1 - \lambda)ln\frac{1}{2}(e^{-i\psi} + 1 + \epsilon) + i\psi + \frac{t}{\lambda} + \widehat{z}(\psi, t) \tag{3}$$

At moment t, the finger tip is located at $\psi = 0, \phi = 0$ and the rear end at $\psi = \pm\pi, \phi = 0$, therefore the length of the finger is

$$L(t) = x_0 - x_1 = 2(1 - \lambda)ln(1 + \frac{2}{\epsilon}) \tag{4}$$

In our notation Bensimon's equation [8] in case of the absence of surface tension is written as

$$\frac{\partial z}{\partial t} = -\zeta\frac{\partial z}{\partial \psi}G(\zeta), \tag{5}$$

where

$$G(\zeta) = A\{g(\psi)\}, \quad g(\psi) = \left[\frac{1}{|\zeta\partial_\zeta z G(\zeta)|^2}l\right]_\Gamma, \tag{6}$$

and the suffix Γ means that the value of the function in the square parentheses is to be taken on the unit circle where $\zeta = e^{-i\psi}$, A is an operator analytically continuing the real function g(ψ) defined on the circumference of the unit circle as the real part of complex function $G(\zeta)$ in the interior of it. The analytical continuation is performed by Fourier expansion.

When t=0, Eq. (2) becomes

$$z = 2(1 - \lambda)ln\frac{1}{2}(\zeta + 1 + \epsilon_0) - ln\zeta \tag{7.1}$$

and the equation of the nonsingular Saffman-Taylor finger is

$$z = 2(1 - \lambda)ln\frac{1}{2}(e^{-i\psi} + 1 + \epsilon_0) + i\psi \tag{7.2}$$

Then

$$g(\psi) = \frac{1 + 2(1 + \epsilon_0)cos\psi + (1 + \epsilon_0)^2}{(1 - 2\lambda)^2 - 2(1 - 2\lambda)(1 + \epsilon_0)cos\psi + (1 + \epsilon_0)^2} \qquad (8)$$

taking a trial solution as

$$G(\zeta) = \frac{a + b\zeta}{(1 - 2\lambda)\zeta - (1 + \epsilon_0)} \qquad (9)$$

with a, b to be decided , we have

$$ReG(\zeta)|_\Gamma = \frac{(b + a)(1 - 2\lambda) - b(1 + \epsilon_0)cos\psi - a(1 + \epsilon_0)}{(1 - 2\lambda)^2 - 2(1 - 2\lambda)(1 + \epsilon_0)cos\psi + (1 + \epsilon_0)^2}. \qquad (10)$$

Equating

$$ReG(\psi)|_\Gamma = g(\psi) \qquad (11)$$

we get

$$b = \frac{(1 - 2\lambda) + (1 - 2\lambda)(1 + \epsilon_0)^2 + 2(1 + \epsilon_0)^2}{(1 - 2\lambda)^2 - (1 + \epsilon_0)^2} \qquad (12)$$

$$a = \frac{(2(1 - 2\lambda) + (1 + \epsilon_0)^2 + 1)(1 + \epsilon_0)}{(1 - 2\lambda)^2 - (1 + \epsilon_0)^2}. \qquad (13)$$

Substituting eqs.(12) and (13) in $G(\zeta)$, we get

$$-\zeta\frac{\partial z}{\partial \zeta}G(\zeta) = \frac{1}{\lambda} - \frac{(1 - \lambda)(1 + \epsilon_0)[1 - (1 + \epsilon_0)^2]}{\lambda[(1 - 2\lambda)^2 - (1 + \epsilon_0)^2](\zeta + 1 + \epsilon_0)} -$$

$$\frac{(1 - 2\lambda)(1 - \lambda)[1 - (1 + \epsilon_0)^2]\zeta}{\lambda[(1 - 2\lambda)^2 - (1 + \epsilon_0)^2](\zeta + 1 + \epsilon_0)}. \qquad (14)$$

Meanwhile, from Eq.(2) at time t, we have

$$\frac{\partial z}{\partial t} = \frac{1}{\lambda} + \frac{2(1 - \lambda)}{\zeta + 1 + \epsilon}\frac{d\epsilon_0}{dt} + \frac{\partial \widehat{z}}{\partial t} \qquad (15)$$

Comparing Eq.(14) and Eq.(15), we have

$$\frac{2(1 - \lambda)}{\zeta + 1 + \epsilon_0}\frac{d\epsilon_0}{dt} = -\frac{(1 - \lambda)(1 + \epsilon_0)[1 - (1 + \epsilon_0)^2]}{\lambda[(1 - 2\lambda)^2 - (1 + \epsilon_0)^2](\zeta + 1 + \epsilon_0)} \qquad (16)$$

$$\frac{\partial \widehat{z}}{\partial t} = -\frac{(1 - 2\lambda)(1 - \lambda)[1 - (1 + \epsilon_0)^2]\zeta}{\lambda[(1 - 2\lambda)^2 - (1 + \epsilon_0)^2](\zeta + 1 + \epsilon_0)} \qquad (17)$$

and at the boundary

$$\frac{\partial \widehat{z}}{\partial t}|_\Gamma = -\frac{(1 - 2\lambda)(1 - \lambda)[1 - (1 + \epsilon_0)^2]\{[1 + (1 + \epsilon_0)cos\psi] - i(1 + \epsilon_0)sin\psi\}}{\lambda[(1 - 2\lambda)^2 - (1 + \epsilon_0)^2][1 + 2(1 + \epsilon_0)cos\psi + (1 + \epsilon_0)^2]}.$$

$$(18)$$

In the first time step we have

$$\Delta\widehat{z} = \frac{\partial \widehat{z}}{\partial t}\Delta t \qquad (19)$$

thus

$$\Delta \widehat{x} = Re \frac{\partial \widehat{z}}{\partial t} \Delta t = -\frac{(1-2\lambda)(1-\lambda)[1-(1+\epsilon_0)^2][1+(1+\epsilon_0)cos\psi]}{\lambda[(1-2\lambda)^2-(1+\epsilon_0)^2][1+2(1+\epsilon_0)cos\psi+(1+\epsilon_0)^2]} \quad (20)$$

$$\Delta \widehat{y} = Im \frac{\partial \widehat{z}}{\partial t} \Delta t \frac{(1-2\lambda)(1-\lambda)[1-(1+\epsilon_0)^2](1+\epsilon_0)sin\psi}{\lambda[(1-2\lambda)^2-(1+\epsilon_0)^2][1+2(1+\epsilon_0)cos\psi+(1+\epsilon_0)^2]} \quad (21)$$

Choosing a suitable ϵ_0 from Eqs.(20) and (21), we can see

- If originally $\lambda > 0.5$, then $\Delta \widehat{x} > 0, \Delta \widehat{y} < 0$, the finger is being stretched to be longer and thinner in later time, i.e., λ tends to decrease.

- If originally $\lambda < 0.5$, then $\Delta \widehat{x} < 0, \Delta \widehat{y} > 0$, the finger is being compressed to be shorter and wider in later time, i.e., λ tends to increase.

- If originally $\lambda = 0.5$, then $\Delta \widehat{x} = 0, \Delta \widehat{y} = 0$, the finger keeps its width forever.

Thus in our one-step dynamic evolution theory about the evolution of nonsingular Saffman-Taylor finger, we conclude: *There exists a unique steady width $\lambda = 0.5$ of the viscous finger to which the fingers of other width approach.* This is exactly what Saffman and Taylor observed in the experiment in 1958.

3 Numerical Computation

From the one-step analysis described above we can not predict completely the shape change in the long run, however, the evolution may be obtained by numerical computation. At time t, if $z(\psi, t) = x(\psi, t) + iy(\psi, t)$ is given, considering $\zeta = e^{-i\psi}$ we have

$$[\zeta \partial_\zeta z]_\Gamma = i\partial_\psi z \quad (22)$$

substitution of eq.(22) in eq.(6) yields

$$g(\psi) = \frac{1}{(\partial_\psi y)^2 + (\partial_\psi x)^2} \quad (23)$$

According to the Poisson integral formula[9], if

$$g(\psi) = a_0 + \sum_{n=1}^{\infty} (a_n e^{in\psi} + a_n^* e^{-in\psi}) \quad (24)$$

we have

$$G(\zeta) = A\{g(\psi)\} = a_0 + 2\sum_{n=1}^{\infty} (a_n \zeta^n) \quad (25)$$

Noticing the symmetry of the finger and substituting Eqs.(22) and (23) into Bensimon equation (5), we get the nonlinear partial differential equation

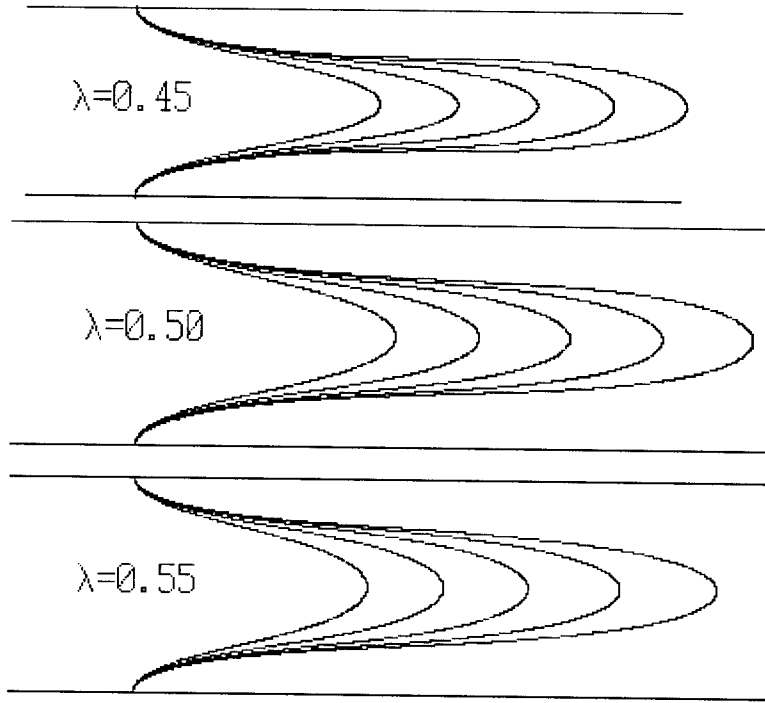$$\partial_t z = -[2\sum_{n=1}^{\infty} (a_n sinn\psi) + ig(\psi)]\partial_\psi z \quad (26)$$

Figure 2. Stable evolution of NST with different $\lambda's, \epsilon_0 = 0.3$

where

$$a_n = \frac{1}{\pi} \int_0^\psi \frac{g(\psi)}{2} cosn\psi d\psi$$

If the initial condition $z(\psi, 0) = x(\psi, 0) + iy(\psi, 0)$ is given, we can compute the equation (26) with various λ and get a family of evolution. Fig. 2 shows the evolution of nonsingular Saffman-Taylor finger. It can be seen from the figure that the above conclusions are valid in the long run of evolution.

4 Conclusions and Prospect

Using the conformal mapping method presented by Bensimon, we present a nonsingular Saffman-Taylor finger. Both theoretical and numerical solutions verify the pattern selection: the tracing finger always tends to $\frac{1}{2}$-width finger in the long run with zero absence surface tension which resolves the Saffman-Taylor's puzzle. Being a simple example of pattern selection which has real experimental background, this work contradicts the generally accepted belief that surface tension is indispensable for the selection of the $\frac{1}{2}$-width finger, a result in parallel to paper in ref [7].
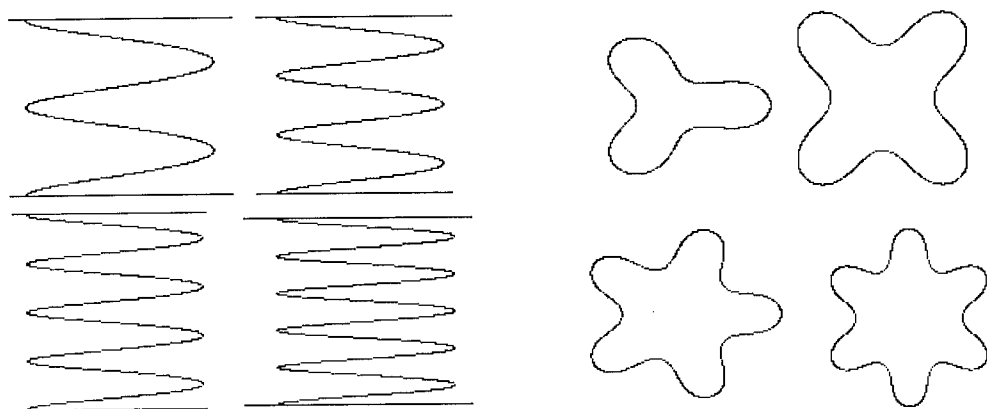
Figure 3. Straight multiple fingers (left) and round fingers (right)

Futhermore, the solution (2) of single finite finger can be generalized to cases of multiple fingers

$$z = \frac{2(1-\lambda)}{n} ln\frac{1}{2}[e^{-n\Phi+i(n-1)\pi} + 1 + \epsilon] + \Phi \qquad (28)$$

and of round fingers

$$lnz = \frac{2(1-\lambda)}{n} ln\frac{1}{2}[e^{-n\Phi+i(n-1)\pi} + 1 + \epsilon] + \Phi \qquad (29)$$

(see Figs. 3 and 4 respectively), may produce more interesting pattern selection problems of evolution, competition and ramification of multiple fingers in straight channel as well as circular disc geometry.

It should be mentioned that Bensimon's equation is a highly nonlinear equation which meets computational difficulty. Tiny round-off errors, especially the residual errors in summing up the Fourier series of Eq.(26) will be magnified quickly. Special efforts has been made to reduce these errors to extremely low level in order to make the computation possible, which has been completed [10]. If surface tension exists, we can compute Bensimon's equation in the same way, but special method of numerical computation must be designed.

Pattern selections of nonsingular Saffman-Taylor finger in the presence of surface tension and multi-fingers constitute our further work.

286

References

1. Saffman P.G.& G.I.Taylor, Proc. Roy. Soc., London Ser. A, 245, 312(1958).
2. H.J.S.Hele-shaw, Nature, 58, 34(1898).
3. McLean J.W.& P.G.Saffman, J. Fluid Mech., 102, 455(1981).
4. M.Vanden-Brock, Phys. Fluids, 26, 2033(1983).
5. Dynamics of curved Fronts, edited by P. Pelce (Academic press, Inc., Boston, 1988).Asymptotic Beyond All Orders, edited by H. Segur et al., NATO ASI, Ser. B. ,Vol 284(Plenum Press, New York, 1991).B.I. Shraiman, Phys. Rev. Lett.,56, 2028(1986).
6. D. C. Hong and J. S. Langer, Phys. Rev. Lett., 56 ,2032(1986); R. Combescot et al., Phys. Rev. Lett., 56 2036(1986). S. Tanveer, Phys. Fluids , 30, 1589(1987).
7. Mark Mineev-Weinstein, Phys. Rev. Lett., 80 , 2113(1998).
8. D.Bensimon, Phys. Rev. A, 33, 1302(1986).
9. L.V.Ahlfors, Complex Analysis (McGraw-Hill, New York, 1953).
10. Huida Yu & Zhen Han, Phys. Rev. E, 58(5), 6873(1998).

A FAMILY OF COMPLEX WAVELETS FOR THE CHARACTERIZATION OF SINGULARITIES

MARIA HAASE

Institut für Computeranwendungen, Universität Stuttgart,
Pfaffenwaldring 27, D-70569 Stuttgart, Germany
E-mail: mh@ica.uni-stuttgart.de

Generally, the irregular behaviour of a function $f(x)$ is described by its Hölder exponent $h(x)$ quantifying the strength of the singularity at a point x. In most cases, where the function does not contain oscillating singularities, the Wavelet Transform Modulus Maxima method (WTMM) allows a reliable estimation[1,2] of the singularity spectrum $D(h)$, i.e. the Hausdorff dimension of the set of all points x with the same Hölder exponent $h(x) = h$. Methods based on the WTMM should be numerically more stable than direct methods such as the Structure Function method[1], since they involve only weighed-averaged quantities and not averages of function increments and since small errors are relatively less important if the calculation is restricted to maxima.

However, in the presence of oscillating singularities the standard WTMM method gives irrelevant information on the Hölder regularity of the function. In general, *two* exponents h, β are necessary to describe the singular behaviour of a function $f(x)$, namely the Hölder exponent h and the oscillation exponent β describing the local power law divergence of the instantaneous frequency[3,4]. If $f(x)$ contains oscillating singularities the regularity of the primitive of $f(x)$ depends on β. In this case the Hölder exponent does not increase by 1 as in the case of a cusp singularity but by $\beta + 1$. Thus, the singularity spectrum of general functions depends on both exponents, $D(h, \beta)$.

In order to extract Hölder exponents and to quantify at the same time the oscillating behaviour we propose to use a family of *complex* progressive wavelets[5] $\{\psi_n\}$

$$\psi_0(x) = (e^{i\omega_0 x} - e^{-\omega_0^2/2})e^{-x^2/2}, \qquad \psi_n(x) = \frac{d}{dx}\psi_{n-1}(x) \qquad (n \in \mathbf{Z}) \qquad (1)$$

with an increasing number of vanishing moments

$$\int_{-\infty}^{+\infty} x^k \psi_n(x)dx = 0 \qquad \text{for} \quad 0 \leq k \leq n \qquad (2)$$

For a sufficiently large frequency ω_0 ($\omega_0 > 5$ for practical purposes), ψ_0 can be replaced by the Morlet wavelet.

Using these wavelets we follow similar ideas as in an earlier paper[2], where an algorithm was developed for a family of real valued wavelets based on derivatives of the Gaussian distribution, which allows a *direct tracing* of the skeleton of wavelet transform maxima lines[2]. Using the following abbreviation for the continuous

wavelet transform

$$W^n f := W^n f(a,b) = \frac{1}{a} \int\limits_{-\infty}^{+\infty} f(x)\overline{\psi_n(\frac{x-b}{a})}dx \qquad (a,b \in \mathbf{R}, a > 0) \qquad (3)$$

where a denotes the scale, b the shift parameter, and $\overline{\psi_n}$ the complex conjugate of ψ_n, we set up a partial differential equation for the wavelet transform

$$\left(a\frac{\partial^2}{\partial b^2} - \frac{\partial}{\partial a} - i\omega_0\frac{\partial}{\partial b} + \frac{n}{a}\right)W^n f = 0 \qquad (4)$$

Describing the maxima line in a parametric form $\{a(t), b(t)\}$, the motion along this line is given by two ordinary differential equations

$$\frac{da}{dt} = -C\frac{\partial}{\partial b}\left(\mathrm{Real}(\frac{\partial W^n f}{\partial b}\overline{W^n f})\right) \quad , \quad \frac{db}{dt} = C\frac{\partial}{\partial a}\left(\mathrm{Real}(\frac{\partial W^n f}{\partial b}\overline{W^n f})\right) \qquad (5)$$

with a constant C, which can be integrated numerically. Similar equations can be derived for the ridges [5]. The skeleton of maxima lines can be continuously traced up to the desired accuracy with a reduced computational effort. Based on these maxima lines a partition function $Z(p,q,a)$ is defined

$$Z(p,q,a) = \sum_{i \in \mathrm{max.lines}} (\sup_{a' < a} |W^n f(a,b_i)|)^q d_i(a)^p \qquad (6)$$

where $d_i(a)$ denotes the distance between the i-th and the $(i+1)$-th maxima line. In comparison with *classical* methods, the main advantage of the proposed direct tracing algorithm is, that the supremum along each maxima line can be easily evaluated.

Exploiting the scaling behaviour of the partition function $Z(p,q,a) \sim a^{\tau(p,q)}$ and applying the Legendre transform to the exponent $\tau(p,q)$, one obtains the two-dimensional $D(h,\beta)$ singularity spectrum[3]

$$D(h,\beta) = \min_{p,q}(qh + p(\beta+1) - \tau(p,q)) \qquad (7)$$

In practice, due to the finite length and noise in data sets, the accuracy of the Legendre transform might decrease. In order to avoid this problem, we propose a generalization of the canonical method of Chhabra and Jensen[6] to two-dimensional spectra which allows a direct computation of the singularity spectrum $D(h,\beta)$.

References

1. J.F. Muzy, E. Bacry and A. Arneodo, *Int.J.Bif. and Chaos* **4**, 245 (1994).
2. M. Haase, B. Lehle, in: *Fractals and Beyond*, (M. M. Novak Ed.), World Scientific, Singapore 241 (1998).
3. A. Arneodo, E. Bacry, S. Jaffard, and J.F. Muzy, *J. Statist. Phys.* **87**, 179 (1997).
4. A. Arneodo, E. Bacry and J.F. Muzy, *Phys.Rev.Lett.* **74**, 4823 (1995).
5. B. Torrésani, in: *Progress in Wavelet Analysis and Applications*, (Y. Meyer and S. Roques Eds.) Editions Frontières (1993).
6. A. Chhabra and J.V. Jensen: *Phys.Rev.Lett.* **62**, 1327 (1989).

Stabilization of chaotic amplitude fluctuations in multimode, intracavity-doubled solid-state lasers [a].

Monika E. Pietrzyk

Institute of Solid State Physics and Theoretical Optics
Friedrich-Schiller-Universität Jena
Max-Wien-Platz 1, 07-743 Jena, Germany
tel: (+49-3641) 947-174, fax: (+49-3641) 947-152
e-mail: p8pimo@rz.uni-jena.de, mpietrz@ippt.gov.pl

Intracavity doubled solid state lasers based on Nd-doped crystals are efficient and compact sources of coherent visible optical radiation. When such lasers operate in three or more longitudinal cavity modes, irregular fluctuations of the output intensity may occur. This behavior, referred to as the green problem, has been reported for the first time by Baer [1]. He found that these instabilities arise from a coupling of the longitudinal modes of the laser by sum-frequency generation, which occur in the intracavity-doubling crystal. When the laser does not contain the nonlinear crystal or when it operates in a single longitudinal mode, its output is stable. In the case of two oscillating longitudinal modes, output intensity of the laser is stable only for small values of nonlinearity, otherwise both modes tend to pulse on and off out of phase with each other [1]. When the number of lasing modes is larger than two, the laser can exhibit, depending on the parameters describing it, various behaviors like: aniphase dynamics [2],[3], clustering [1], grouping [4] and chaotic dynamics [5].

The main goal of this communication is to study the possibilities of a stabilization of large amplitude fluctuations in such a laser, i.e., an intracavity-doubled Nd:YAG laser. The analysis is based on the Baer-type rate equations [1]

$$\tau_{c1}\frac{\partial I(p,t)}{\partial t} = I(p,t)\left(-\alpha_p + G(p,t) - \epsilon I(p,t) - 2\epsilon\sum_{q\neq p} I(q,t)\right), \qquad (1a)$$

$$\tau_f\frac{\partial G(p,t)}{\partial t} = G_{ap} - G(p,t)\left(1 + \beta(p,p)I(p,t) + \sum_{q\neq p}\beta(p,q)I(q,t)\right), \qquad (1b)$$

$$p, q = 1, ..., N,$$

where N is the number of longitudinal modes; τ_c and τ_f are the cavity round trip time and the fluorescence lifetime of the Nd^{+3} ion, respectively; $I(p,t)$ and $G(p,t)$ are, respectively, the intensity and gain associated with the p-th longitudinal mode; α_p is the cavity loss parameter for the p-th mode; G_{ap} is the small signal gain; $\beta(p,p)$ is the self-saturation coefficient in the active medium; $\beta(p,q)$ is the parameter describing the cross-saturation between two longitudinal modes, p and q. The parameter ϵ is a nonlinear coefficient whose value depends on properties of the nonlinear crystal and it describes the conversion efficiency of the fundamental

---

[a]**Key words**: Chaos in laser cavities; Multimode, intracavity-doubled solid-state lasers; Baer-type rate equations

intensity into the doubled intensity. The terms $\epsilon I(p,t)^2$ and $\epsilon I(p,t)I(q,t)$ in Eq. (1a) account for the loss in the intensity of the fundamental frequency through second harmonic generation and through sum-frequency generation, respectively, and they provides a nonlinear loss mechanism that globally couples the longitudinal modes, i.e. each lasing mode is coupled to all other lasing modes [2]. A comparable amount of global coupling occurring in the set of Eq. (1) is introduced by the cross-saturation coefficient, $\beta I_k G_k$. We use the approximation that the cross-saturation coefficient is constant for all modes, $\beta(p,q) = \bar{\beta}$, where $\bar{\beta} = \frac{2}{3}\beta_0$ or $\bar{\beta} = \frac{1}{3}\beta_0$, where $\beta_0 = 0.06$ is a scalling parameter. We assume also that the losses and the small signal gains are the same for all modes, i.e. $\alpha_p = \alpha_q = \alpha$, $G_{ap} = G_{aq} = G_a$, where $p,q = 1, ..., N$. Other parameters describing the system we have chosen in such a way that they can describe a real experimental configuration of the laser, i.e., $\tau_{c1} = 10\,[ns]$, $\tau_f = 0.24[ms]$, $\alpha = 0.015$, $\gamma = 0.12$. The number of longitudinal modes, $N = 1, ..., 250$, and the strength of nonlinearity, $\epsilon = 10^{-7} \div 10^{-3}$, are not fixed and they vary in the analysis.

First, we analyze numerically the stabilization of the laser radiation by an increase of the number of longitudinal modes, proposed in [2]. We observe that the theoretically obtained [3] linear dependence of the minimal number of modes, which are needed to stabilize the laser output, on the strength of nonlinearity agree with the numerical solutions only in the case of sufficiently small nonlinearity. For large nonlinearity the minimal number of modes obtained by the numerical simulations is larger than the number which follows from the theoretical predictions. It is caused by a strong cancellation of modes during the evolution. For very large nonlinearity this cancellation is so strong that only few modes remain (even when there are initial 250 oscillating modes). Therefore, a large number of simultaneously oscillating longitudinal modes cannot be achieved in this case.

However, the problem of the stabilization of the laser output can be solved in another way, namely, by an increase of the strength of nonlinearity, which leads to very strong competition between the modes, so that during the evolution all of them, besides a single one, are canceled. As a consequence, a steady-state solution, which is stable against small perturbations, arises.

This way of stabilization, achieved by forcing the laser to operate in one-mode regime, is similar to other approaches presented in the literature, where the stabilization is obtained by inserting into the laser cavity an additional element like, for example, an etalon [1] or a birefringent crystal. However, this proposed method seems to be a better solution, since no additional element is needed and the output intensity of the doubled frequency is larger (it increases with the increasing of the strength of nonlinearity).

References

1. T. Baer, J. Opt. Soc. Am. B **3**, 1175 (1986).
2. G. E. James *et al.*, Opt. Lett. **15**, 1141 (1990).
3. J.-Y. Wang and P. Mandel, Phys. Rev. A **48**, 671 (1993).
4. K. Otsuka, Y. sato, and J. L. Chern, Phys. Rev. E **56**, 4765 (1997).
5. M. E. Pietrzyk and M. B. Danailov, preprint of the ICTP (the Abdus Salam International Centre for Theoretical Physics), Trieste, Italy, **IC/98/235** .

CHAOTIC DYNAMICS OF ELASTIC-PLASTIC BEAMS

ÜLO LEPIK

Institute of Applied Mathematics, Tartu University, Vanemuise 46, 51014 Tartu, Estonia
Phone: (3727)375868, Fax: (3727)375862, E-mail: ylepik@ut.ee

During the last decades great attention has been turned to chaotic vibrations of elastic beams, but not much has been done in the case of elastic-plastic deformations. Evidently the first paper in this field belongs to Symonds and Yu (1985), who considered the following problem. A fixed ended beam is subjected to short intensive pulse of transverse loading that produces plastic deformation. Since the ends of the beam are fixed membrane forces must be taken into account. Solving the equations of motion Symonds and Yu found that permanent deflection may be in direction opposite to the load. This phenomenon was investigated in several papers by Symonds and his collaborators. It turned out that the permanent defflection is very sensitive to small changes of load. Fractality and self-similarity which are characteristic to chaotic processes, were demonstrated. For the similarity dimension the value 0,78 and for the correlation fractal dimension ~1,44 were obtained.

In most papers of Symonds and his collaborators the real beam was replaced by a Shanley-type model with one or two degrees of freedom. In some papers (see Symonds and Qian, 1996) the Galerkin method for $2DoF$ was used; some simplifying assumptions (sandwich beam, perfectly plastic material, disregard of axial inertia forces) were made.

The aim of this paper is to present a method of solution, which is based on the Galerkin technique and is applicable for beams with an arbitrary number of DoF. Basic elements of this method were worked out by Lepik (1994, 1995).

To be brief we shall consider beams with a rectangular cross-section, B, h and L are the width, thickness and length of the beam, respectively. The equations of motion are

$$\Phi_1 = \frac{\partial T}{\partial x} - \rho Bh \frac{\partial^2 u}{\partial t^2} = 0,$$
$$\Phi_2 = \frac{\partial^2 M}{\partial x^2} + \frac{\partial}{\partial x}\left(T\frac{\partial w}{\partial x}\right) - p(x,t) - \rho Bh \frac{\partial^2 w}{\partial t^2} = 0,$$

(1)

where ρ is density, u - axial displacement, w - deflection. Axial force T and bending moment M we shall calculate from the formulae

$$T = \int_{-h/2}^{h/2} \sigma(x,z)\, dz, \quad M = \int_{-h/2}^{h/2} \sigma(x,z)z\, dz.$$

(2)

We shall assume that the beam material has linear strain-hardening; elastic unloading and secondary plastic loading effects are also taken into account.

To the equations (1) we shall apply the Galerkin procedure

$$\int_0^L \Phi_1 \delta u\, dx = 0, \quad \int_0^L \Phi_2 \delta w\, dx = 0.$$

(3)

In the case of a beam with simply supported ends we shall seek the solution in the form

292

$$u = e_0 \left(1 - \frac{x}{L}\right) + \sum_{k=1}^{s} a_k \sin k\pi \frac{x}{L}, \qquad w = \sum_{k=1}^{s} f_k \sin k\pi \frac{x}{L}, \qquad (4)$$

where $e(0) = e_0$ and the coefficients e_0, a_k, f_k are subject to variation. The integrals (2) will be calculated numerically; replacing these results into equations (3) we find the second derivatives $\ddot{e}_0, \ddot{a}_k, \ddot{f}_k$. The quantities e_0, a_k, f_k are evaluated according to the method of central finite differences (for getting a stable solution the time increment Δt must be sufficiently small).

In order to demonstrate the efficiency of the proposed algorithm several numerical examples are examined. With the purpose to establish chaotic effects the deflection history, phase and power spectrum diagrams are put together. It turned out that chaotic motion of the beam may exist, especially in the initial phase of motion; as to the long term motion then it transitates to periodic vibrations of smaller amplitude.

References

1. Lepik Ü., Dynamic response of elastic-plastic pin-ended beams by Galerkin's method. *Int. J. Solids Structures*, **31**, (1994), 3249-3260.
2. Lepik Ü., Elastic-plastic vibrations of a buckled beam. *Int. J. Non-Linear Mech.*, **30**, (1995), 129-139.
3. Symonds P.S. and Yu T.X., Counterintuitive behavior in a problem of elastic-plastic beam dynamics. *ASME J. Appl. Mech.*, **52**, (1985), 517-522.
4. Symonds P.S. and Qian Y., Aspects of chaotic response of an elastic-plastic beam to impulsive loading. Proceedings of the IUTAM and ISIMM Symposium on Anisotropy, Inhomogencity and Nonlinearity in Solid Mechanics (University of Nottingham, 1994), Kluwer, 1995.

THE RIEMANN NON-DIFFERENTIABLE FUNCTION AND IDENTITIES FOR THE GAUSSIAN SUMS

B. T. MATKARIMOV

Institute of Mathematics, Pushkin str. 125, Almaty 480100, Kazakstan
E-mail: matkarimov@itte.kz

Riemann's example of a continuous, non-differentiable function is given by the sum $\sum_{n=1}^{\infty} \sin(n^2 x)/n^2$. This function is sufficiently irregular and its graph is fractal. Hardy proved [1] that Riemann's non-differentiable function is not differentiable at any irrational point because of a square root singularity at these points. Careful investigation of the differentiability of Riemann's non-differentiable function was carried out by Gerver, who showed [2] that this function has derivative equal to $-1/2$ at every rational point of a special type (forming the orbit of the point 1 under the theta-modular group[3]). Different proofs of this surprising fact was given by other authors[3,4,5], providing also a close relation between Riemann's non-differentiable function and classical θ-function and Gauss sums. Duistermaat obtained [3] an exact functional equations for this function under transformations of the theta-modular group. In this article we use functional equations for Riemann's non-differentiable function under theta-modular transformations to derive functional equations on Gauss sums generalizing Genocci-Schaar identity.

A Gauss sum is a sum of the form

$$S(p, q) = \sum_{n=0}^{q-1} \exp(\pi i n^2 p/q),$$

where p and q are relatively prime integers of opposite parity, i.e. one is odd and the other is even. The Genocci-Schaar identity on Gauss sums is the following: for positive integers p and q of opposite parity,

$$\frac{1}{\sqrt{q}} \sum_{n=0}^{q-1} \exp(\pi i n^2 p/q) = \frac{\exp(\pi i/4)}{\sqrt{p}} \sum_{n=0}^{p-1} \exp(-\pi i n^2 q/p).$$

This identity can be interpreted as the transformation of the Gauss sum under the change $\sigma : z \to -1/z$ where $z = p/q$.

The modular group Γ is a group of fractional linear transformations $\gamma : \ z \to (az + b)/(cz + d)$ where $a, b, c, d \in Z$ and $ad - bc = 1$. Theta-modular group Γ_θ is a sub-group of modular group generated by the following mappings: $\tau : z \to z + 2$ and $\sigma : z \to -1/z$. For any element of the theta-modular group the following are valid: $ab = 0 \ (mod \ 2)$ and $cd = 0 \ (mod \ 2)$. Every fractional transformation $\gamma \in \Gamma_\theta$ has a simple pole at the point $z = -d/c$. The rational points $x = p/q$ with integers p, q of opposite parity and infinity constitute theta-modular orbit of point 0.

Riemann's non-differentiable function has the following local estimations at the point x: $f(x \pm \epsilon) = f(x) + R_\pm \epsilon^\delta$, where $0 < \delta \leq 1$. Using the functional equations for $f(x)$ it is possible to find functional equations on the functions $R_\pm$. These functions are known only at the rational points, where they coincide with the Gauss sums. The following theorems describe these functional equations:

Theorem: Let r and s be integers of which one is even and the other is odd, and s is positive, $\gamma \in \Gamma_\theta$ be an element of the theta-modular group $\gamma : z \to (az + b)/(cz + d)$, and suppose c is positive and $cr + ds \neq 0$. Define $r' = ar + bs$ and $s' = cr + ds$. Then the following formula for Gauss sums is valid:

$$\frac{\exp(\pi i Sign(s')/4)}{\sqrt{|s'|}} S(r', s') \frac{1}{\sqrt{c}} S(-d, c) = \frac{1}{\sqrt{s}} S(r, s) \qquad (1)$$

Proof: The proof is based on the differentiability properties of Riemann's non-differentiable function. Let us consider the following function $\phi(z) = \sum_{n=1}^{\infty} \exp(\pi i n^2 z)/\pi i n^2$. Using technique of papers[4,5] we calculate the following estimation for the function $\phi(z)$ at the point $x = u/v$, where u and v are relatively prime integers of opposite parity, $(u, v) = 1$ and $uv = 0 \ (mod\ 2)$:

$$\phi(x + h^2) - \phi(x - h^2) = h^2 \sum_{n=-\infty}^{+\infty} \exp(\pi i n^2 x)\varphi(nh) - h^2 = \qquad (2)$$

$$h^2 \sum_{t=0}^{|v|-1} \exp(\pi i t^2 x) \sum_{k=-\infty}^{+\infty} \varphi(k|v|h + th) - h^2 = 2^{1/2}S(u, v)h/|v| + O(h^2).$$

Here $\varphi(x) = \sin(\pi x^2)/\pi x^2$ if $x \neq 0$ and $\varphi(x) = 1$ if $x = 0$, and we write $n = k|v| + t$, $(0 \leq t \leq |v|)$ and use that $\exp(\pi i n^2 x) = \exp(\pi i t^2 x)$, since $uv = 0 \ (mod\ 2)$. The function $\phi(z)$ obey a functional equation under the action of theta-modular group[3]. Let $\gamma \in \Gamma_\theta$ be an element of theta modular group, then the function $\psi(z) = \phi(z) - \gamma'(z)^{-1}\mu_\gamma(z)\phi(\gamma(z))$ is differentiable, and analytical function μ_γ is given by:

$$\mu_\gamma(z) = e^{(\pi i/4)}c^{-1}(z + d/c)^{-1/2}S(-d, c). \qquad (3)$$

Eq. (2) is valid for any point of theta-modular orbit of $x = r/s$, except infinity. Calculating estimation (2) for the differentiable function $\psi(x)$ at the point $x = r/s$ and supposing $\gamma(r/s) \neq \infty$, we find that

$$\frac{S(r, s)}{s} - \gamma'(r/s)^{-1/2}\mu_\gamma(r/s)\frac{S(ar + bs, cr + ds)}{|cr + ds|} = 0. \qquad (4)$$

Note, that $r' = ar + bs$ and $s' = cr + ds$ are relatively prime integers. Substituting the expression of $\mu_\gamma(z)$ in terms of Gauss sums (Eq. 3) we obtain the desired result.
Q.E.D

The Genocci-Schaar identity corresponds to the case of $a = 0$, $b = -1$, $c = 1$ and $d = 0$. Functional equation (Eq. 1) describe the theta-modular transformations of Gauss sums and can be used to derive the values of the Gauss sums.

References

1. G. H. Hardy, *Trans. AMS* **17**, 301 (1916)
2. J. Gerver, *Amer. J. Math.* **92**, 33 (1970)
3. J. J. Duistermaat , *Nieuw Archief voor Wiskunde* **9**, 303 (1991)
4. A. Smith, *Proc. AMS* **34**, 463 (1972)
5. W. Luther, *J. Approx. Theory* **48**, 303 (1986)

REVEALING THE MULTIFRACTAL NATURE OF FAILURE SEQUENCE

R. I. ZAINETDINOV

Moscow State University of Railway Communication, 15 Obraztsov Street, Moscow, 101475, Russia
E-mail: zri@hotmail.com

The purpose of the paper is to introduce a new application of the multifractals in the reliability engineering, risk analysis, historical chronology and others fields where we deal with series of events of various natures. The spectrum of such series is quite broad. In particular, we do not know whether some temporal pattern is hidden in an apparently disordered set of events. The multifractal theory is a good basis for revealing such an order and describing the event sequence in time. It can provide a deeper understanding the nature of the event flow.

A mathematical construction that represents an event sequence as a set of the random points on the time scale is referred to as a stochastic point process. It can either be modeled as a list of impulses located at times where events occur or as a count process. Let S be a sample of events of limited size N_0 during the specified period of time $[0, \tau_{max}]$. The process time history for the sample S is represented by a sequence of idealized impulses of vanishing width, located at specified moments of the event time τ_i, $i = 1, 2,..., N_0$. Further, let us rescale the time τ_i of every i-th member of sample S on the maximum value τ_{max} $t_i = \tau_i/\tau_{max}$, so we can consider the event distribution on the unit interval of time $T = [0, 1]$. In order to characterize this distribution we divide the unit interval into temporal subintervals of duration $\Delta t = 2^{-n}$. So $N = 2^n$ subintervals are needed to cover interval T, where n is the number of generation in the binary subdivision of the temporal interval T. The distribution of the sample population over the temporal interval is specified by the numbers, N_j, of members of the sample S in the j-th subinterval. We use the fraction of the total population $\mu_j = N_j/N_0$ as a probabilistic measure for the content in subinterval Δt_j. The set of such measures presents a complete description of the event's distribution on the unit temporal interval T at stated resolution Δt [3].

Now let us consider a case that satisfies the Bernoulli trial conditions. In our interpretation an event of interest is the failure occurred on the first half of temporal interval with probability p. The series of Bernoulli trials with parameter p is a sequence of independent trials in which there are only two outcomes, and probability p remains the same for all generations of the binary subdivision process of interval T. Three cases associated with typical form of cumulative distribution function are considered: process with early events (when $p>0.5$), Poisson process ($p=0.5$), and process with late events ($p<0.5$). In the case of Bernoulli trials the measure μ is recursively generating by a multiplicative binomial process. The binomial cascade provides an example of a probability, which has a rich asymptotic structure and is, in modern terms, multifractal [3]. Connection between parameter p of the Bernoulli trial and multifractal spectrum is considered.

For the approbation of the technique, a computer simulation study has been done. As a first step, we have carried out a multiscale analysis of data generated by Bernoulli trials. The binomial multiplicative process produces shorter and shorter temporal subintervals Δt that contain less and less fractions of the total measure and generates a multifractal probabilistic measure, supported by Cantor set on the unit temporal interval.

In order to verify the fractality of data sets obtained from numerical simulations, tests and inspections of real devices a wavelet analysis was carried out. The analysis shows

that maxima lines of the continuous wavelet transform (CWT) coefficients are converging towards the singularities of the measure, and they reproduce its hierarchical structure [1]. The successive forkings occurred at different scales reveal the multifractal nature of measure. Failure occurrence is probabilistic process, which results in the formation of self-similar, or rather self-affine temporal clusters. Overwhelming evidence from computer simulation of different measures indicates that these patterns are self-affine fractals, meaning that their complication is the same at different scales of observation.

The wavelet analysis was carried out by using the WaveLab package [2].[a] The multifractal structures proposed in the sequence of events are real-time structures, in contrast to fractal attractors, which reside in phase space. Thus the wavelet transform can be applied directly to a series of statistical data on reliability obtained from experiments and inspections of technical state under real service conditions [4]. The graph of the CWT coefficients of the failure time history shows that the successive forkings produce a multifractal temporal structure. Increasing the magnification of the wavelet transform microscope reveals progressively the successive generations of branching. The symmetry of the graph is broken by non-uniformity of probabilistic measure. Let $N(a)$ be the number of maxima lines in the CWT skeleton at the scale a. In the limit, as the scale a tends to 0, the ratio $\ln(N(a)/\ln(a)$ is associated to the exponent α [1]. The concentration of data points around the straight line observed in the plot of $\ln(N(a))$ versus $\ln(a)$ can be regarded as a quantitative indication of the self-similarity of the event sequence in real data sets. The ratio between the time scales of successive generation can take different values, which is another indication that a multifractal description is appropriate. Wavelet analysis of empirical data on reliability provides probabilistic evidence for the existence of a multiplicative process hidden in the temporal ordering of the failure sequence.

The plots of empirical cumulative distribution function of the lifetime data have an evident feature, namely, they are constant almost everywhere except in those points where failures occur. In the limit, the empirical distribution function resembles in a sense a devil staircase prefractal. In the proposed multifractal approach the emphasis switches to letting the statistical data "speak for themselves", rather than approximating the lifetime distribution by one of the parametric models. In practice, we are often interested in prediction of the mean time of a failure-free operation as well as the others reliability indexes. Deeper insight into failures, their prediction and prevention is to be gained by using the multifractal approach in the reliability engineering.

References

1. Arneodo A., Wavelet analysis of fractals: from the mathematical concepts to experimental reality. In *Wavelets. Theory and Applications*, ed. by G. Erlebacher, Y. Hussaini, and L. Jameson (Oxford University Press, New York, 1996) pp. 349-502.
2. Buckheit J., and Donoho D., WaveLab and Reproducible Research. In *Wavelets and Statistics*, ed. by A. Antoniadis and G. Oppenheim (Springer, New York, 1995) pp. 55-81.
3. Feder J., *Fractals* (Plenum Press, New York, 1989).
4. Zainetdinov R. I., Wavelet Analysis of Statistical Data on Reliability for Exploring the Multifractal Properties of Failure-Cascading Process. In *Wavelets and Multiscale Methods* (INRIA, Paris, Tangier, 1998).

THE FRACTAL NATURE OF WOOD REVEALED BY DRYING

BENHUA FEI

Research Institute of Wood Industry, Chinese Academy of Forestry,
Beijing 100091, China

Nature provides us with a considerable variety of shapes and forms, from the simple to the complex. There is much current interest in the application of fractal geometry [1], to describe the complex phenomena related to the natural shapes and the structure exhibiting an intricate and irregular form. The degree of geometric irregularity of an object can be measured by fractal dimension, such as length of coastlines, molecular surface, aerogels, bacterial colonies, tumor cells, axon terminals, tree crown properties, and so on. Wood is a typical porous material that exhibits interconnected pores composed of biological elements. The pore space consists chiefly of cell cavity, intercellular hole and void in cell wall, which shape a complex structural space of wood. Many reports had shown that the pore space characteristics within wood bear a close relation to the water content, because the structure and size of the pore space changes when water evaporates and escapes from wood. There are also some results on the relationships between the pore space and the water content within wood. Most of them, however, did not take into account the objective regularity and can not reveal the complexity of wood structure.

Introducing the fractal theory into this field will provide new insights. The fractal dimension of wood surface was first studied by using water-sorption. It was found [2] that the fractal dimension was in the range of 2.5 — 2.8. An experiment [3] with pressurized water absorption has confirmed the fractal nature of wood. The findings[3] have shown that the pore space within wood could be characterized by fractal dimension or a set of fractal dimensions. But there was no explanation leading to the characterization of wood properties.

A typical property of fractals relates their volume V to the corresponding linear size L, as

$$V(L) \propto L^{df}, \tag{1}$$

where df is in general a non-integral and $df \le d$. Herein d is the Euclidean dimension of the space in which the object is embedded [4]. This relation is familiar to us when dealing with the usual objects such as lines, discs, or spheres. The mass change (ΔM) of a sample, or the mass of bound water evaporated at different temperatures (T), is a function of the linear dimension (L) of the sample and the drying temperature ($\Delta M \equiv \Delta M(L, T)$). As previously stated, the mass change ΔM represents the volume of the pore space in the cell wall of the sample. Hence from Eq. (1), it can be expressed as

$$\Delta M(L, T) \propto L^{df(P)} \tag{2}$$

where $df(P)$ is the fractal dimension of the pore space in cell walls of a wood sample. The constant of proportionality is not important to us; only the scaling is important to the fractal nature of the structure of pore space. Take the natural

298

Table 1. The fractal dimensions of Ginkgo and Chinese chestnut obtained at varable temperatures

species	$20°C$	$40°C$	$60°C$	$80°C$	$100°C$
Ginkgo	2.106	2.547	2.851	2.863	2.876
Chinese chestnut	2.008	2.566	2.814	2.896	2.972

logarithm of both sides of Eq. (2) and we have

$$ln(\Delta M) \propto df(P)ln(L).$$

If a set of ΔM and L is acquired, $df(P)$ can be obtained from a $log - log$ plot of the mass change ΔM and the linear dimension L.

The materials for this investigation came from two species, one was a 37-year-old plantation-grown Ginkgo (Ginkgo biloba) and the other was a 48-year-old plantation-grown Chinese chestnut (Castanea mollissima). After the discs were completely air-dried, they were sawn from pith to bark in the radial direction. Then 11 cubic blocks were sawn near the center of radial face of the discs and sanded. The three dimensions of each cubic block increased in steps of 5mm from 5 mm to 55 mm. The initial moisture of the samples was 14%. The samples were weighed and measured first. Then they were dried in oven at temperatures of $20°C$, $40°C$, $60°C$, $80°C$, and $100°C$, respectively. The duration of drying was four hours in all cases. After each drying was completed, samples were weighted and measured again.

It was found that the fractal dimensions of the pore space of wood increase as the temperature increases. The fractal dimension is an intrinsic property of the pore space and can be as a new parameter for the characterization of wood. The fractal dimensions obtained in this paper are in the range of 2.1057 — 2.8757 and 2.0080 — 2.9718 for Ginkgo and Chinese chestnut, respectively. They reflect the complexity and irregularity of structure of both woods. The values obtained here for the mass changes $\Delta M(L, T)$ are very sensitive to the nature of the volume of the pore space, and the fractal dimensions should depend on the nature of the escaped water used in the experimental procedure. Compared with the experiment with water at a given pressure, the experiment procedure in this paper is simpler and easier to control. Using this method, the geometric set composed of pores, voids, and micro voids in wood can be described.

References

1. Mandelbrot B B. 1977. The fractal geometry of nature, W.H. Freeman and Co., New York, NY
2. Hatzikiriakos S G, S Avramidis. 1994. Fractal dimension of wood surfaces from sorption isotherms, Wood Science and Technology, 28:275-284
3. Jose Arnaldo Redinz and Paulo Roberto Colares Guimaraes. 1997. The fractal nature of wood revealed by water absorption. Wood and Fiber Science. 29 (4):333-339
4. J Zhang, 1997, Fractals. Publishing House of Qinhua University

SQUARING THE CIRCLE : DIFFUSION VOLUME AND ACOUSTIC BEHAVIOUR OF A FRACTAL STRUCTURE

PHILIPPE WOLOSZYN

CERMA Laboratory, UMR CNRS 1563, Ecole d'Architecture de Nantes, rue Massenet,
F-44300 Nantes, France.
E-mail : woloszyn@cerma.archi.fr

The main topic presented in this paper concerns the reduction of the diffusion volume to convex isotropic Euclidean structures as polygons and convex polyhedrons, in order to estimate the diffusive acoustic behavior of a fractal structure.

The volume V and the surface S of an Euclidean structure are defined with both quadratic and cubic relationships to the perimeter P with $S = F_q P^2$, and $V = F_C P^3$. Their geometrical nature depends on the values of the *prefactors* in those relationships, the *shape factors* F_q and F_C [1]. Extension of this relationship to the structure's generalized volume V_x [2] gives us the equation :

$$V_x = F_x P_x^{D}.$$ (1)

Calculation of the corresponding shape factors respectively gives $1/4\pi$ for a circular and $3/32\pi$ for a spherical structure. In order to generalize this expression to the whole platonician 2-D family, we operate a polygonal circle quadrature with the convexes from square to n-sided polygons, $n \to \infty$.

The quadrature factor Q represents the ratio between the circumscribed and inscribed circles of the structural polygons of the analyzed objects. They are varying from $Q = 2$ for the square (minimal quadrature) to $Q = 1$ for the circle (total quadrature). The general expression is given by :

$$Q = \frac{S_{Cir}}{S_{Ins}} = \frac{\frac{R_{Cir}^2}{2} \sin\frac{2\pi}{n}}{r_{Ins}^2 \tan\frac{\pi}{n}}$$ (2)

Taking into account D, fractal dimension and d, Euclidean dimension of the structure, the F_x factor expression becomes :

$$F_x = \frac{d}{2^{dQ}(d-1)2\pi^{|2-Q|}\frac{d^{Q-1}}{Q^{D-1}}}$$ (3)

Eq. (3) gives the F_x value of the square ($F_x = 1/16$), the cube ($F_x = 1/64$), the circle ($F_x = 4\pi$) and the sphere ($F_x = 3/32\pi$). It gives also the intermediate values of the regular polygonals which square the circle, noticing that for dodecagon, the prefactor F_x reaches the circle constant value $F_x = 1/4\pi$. This confirms the quadrature convergence.

By replacing F_x with its expression in equation (3), one can write the perimeter-surface relationship (P / S_x) for irregular 2-D shapes and to closed 2-D fractals, with its quadrature factor Q and its Euclidean dimension d as follows :

$$S_x = \frac{d}{2^{dQ}(d-1)2\pi^{|2-Q|}\dfrac{d^{Q-1}}{Q^{D-1}}} P^D \tag{4}$$

We are using the Minkowski analysis method to scrutinize the structure with perimeter P and real diffusion section S_x. The structuring element of radius Λ we use, gives the fractal dimension of the diffusion volume depending on perimeter P, true diffusion section S_x and the structure's fractal dimension D as follows [3] :

$$D = 1 - \frac{S_x}{P\Lambda} = 1 - \frac{\upsilon\, S_x}{P} \tag{5}$$

This leads to the definition domain of the *diffusion coefficient* δ which evolves [4]:

$$\delta = \frac{1}{2\upsilon\tau} = \frac{S_x(1-D)}{2P\Delta t} \tag{6}$$

where Δt represents the diffusion time-share in the structure. This leads to the diffusion time expression, depending on the fundamental length of acoustic diffusion process λ_{max}, and the roughness parameter n (number of indentations per length unit) expressed as :

$$\tau = \lambda_{max}^{-D} f^{d-1} = \left(F_x(1-D)\frac{P^{d(1+D-d)-1}}{n\Lambda^2} \right)^{-D} f^{d-1} \tag{7}$$

by replacing diffusion time τ with its expression in (7), the diffusion coefficient becomes :

$$\delta = \frac{1}{2\upsilon}\left(F_x(1-D)\frac{P^{d(1+D-d)-1}}{n\Lambda^2} \right)^{D} \frac{1}{f^{d-1}} \tag{8}$$

As we can see equation (8), this coefficient is defined with only the structural morphologic diffusion parameters and the sound frequency f.

That result leads us to confirm that the diffusion coefficient is a geometrical-dependant parameter, confirming the rule of the structure complexity on acoustics.

Thus, after a multiscale renormalisation, the diffusion coefficient has to be spectrally defined by adjusting a simulation procedure to experimental results, in order to implement this diffusive acoustic model for any complex configuration : those perspectives constitutes the next step of our research works.

References

1. Mandelbrot B. : " Les objets fractals ", Paris, Flammarion, 1992
2. Pfeiffer P., Obert M. & Cole M.W. : " Fractal BET and FHH theories of adsorption: a comparative study ", Proc. R. Soc. Lond. A423, pp. 169-188, 1989
3. Woloszyn P. : " Mesures multiéchelles du tissu urbain et paramétrage d'un modèle de diffusion acoustique en milieu construit ", *Symposium Saint-Venant* : *"Analyse multiéchelle et systèmes physiques couplés "*, Marne-la-Vallée, août 1997
4. Chandrasekhar : " Stochastic Problems in Physics and Astronomy ", Review of Modern Physics, 15 (1) : 1-89, 1943

RELATIONSHIP BETWEEN ACUPUNCTURE HOLOGRAPHIC UNITS AND FETUS DEVELOPMENT; FRACTAL FEATURES OF TWO ACUPUNCTURE HOLOGRAPHIC UNIT SYSTEMS

Y. HUANG

The Medical Dept. of the Science & Technology Committee, Shanghai Bureau of Astronauts, Shanghai Delphi Automotive Air Conditioning Systems Co. Ltd., 1768 Hunan Rd., Pudong New Area, Shanghai, China 201204

Chinese acupuncture-moxibustion is a therapeutic method for injecting or burning moxa (a kind of herb) on the acupuncture points (acupoints) of the body surface, used to treat diseases. There are over one thousand acupoints on the body surface. In the clinic, the acupoint system on the human body can be divided into two large parts: the channel system (macro-acupuncture) and the micro-acupuncture system. But the whole acupoint system can be divided into many different hierarchical holographic units. Such system possesses fractal structure [1].

We found that the acupuncture holographic units can be divided into two kinds. In one, their acupoints distribution graphs are similar to the shape of the human nerve-embryo; in the other these graphs are similar to the shape of a mature fetus.

In a 7-week old human fetus (nerve embryo 1.7 cm, 2g) its organ differentiation has already begun, as has its limb growth. Its 'body segment' development is very good. The whole nerve-embryo can be divided into many 'body segments'. Each of these contains a 'nerve segment' a dermatotome of skin (includes acupoints) and the corresponding organ. The nerve segments can control skin and organ in the same body segment.

The whole channel system is an acupuncture holographic unit. The course of the channels runs parallel with spinal column on the trunk. The acupoints in the channel are distributed at equal distances. The channel acupoints distribution possesses segmental structure. It has been found, in the anatomy results, that there is a good relationship between the acupoints distribution on the body surface and that of endings of the spinal nerves. The channel acupoints receive segmental innervation of spinal nerve. Although the extremities have no segmental structure, the channel acupoints on the extremity parts still receive segmental innervation of the spinal nerve. Therefore, the acupoints distribution and innervationn of the channel system retains the segmental structure. The channel system is an acupuncture holographic unit which is similar to the nerve-embryo.

The back acupoint system on the trunk is part of the channel system. It is an acupuncture holographic unit too. The acupoints distribution retains the segmental structure. Its acupoints receive segmental innervation of the spinal nerve. Thus, the back acupoint system is similar to the nerve-embryo.

There is the same micro-acupuncture system in each different hierarchical part of the human body [2]. The acupoints distribution graphs of these micro-acupuncture systems retain segmental structure. They are similar to the shape of the nerve-embryo. But the acupoints in these systems receive super-asegmental innervation.

On the other hand, in some acupuncture units, their acupoints distribution graphs resemble the shape of a fetus. It was found [3] that the acupoints distribution

graph of ear is just like an inverted fetus in a womb. The acupoints of organs originated from the same germinal layer (entoderm, ectoderm and mesoderm) are distributed over the same area in the ear. Te acupoints in the ear system receive superasegmental innervation of the nerve system.

The acupoints distribution graphs of the face- and nose-acupuncture systems are similar to the mature fetus, but there are deformations. These are copies of the shape of the mature fetus that are compressed along different direction with different proportion. The acupoints of face-and nose-acupuncture systems receive super-asegmental innervation.

There are many different hierarchical acupuncture holographic units of the channels on the extremity. Their acupoints distribution graphs are similar to the shape of mature fetus with deformation.

To sum up, the acupuncture holographic units can be divided into two groups: the acupoints distribution graphs either show similarity in shape to the nerve embryo or to a mature fetus.

It is well known that a self-affine system can be divided into many parts, each part is a copy of the whole which is compressed along different directions with different proportions. We have found that this concept can be used to describe the acupuncture holograpic unit system made of units similar to the shape of mature fetus. Each unit in this system is a copy of the mature fetus, which develops along different direction with different proportions. Their acupoint distribution graphs are similar to the shape of mature fetus but with deformations.

The tructure of a self-affine system generator T is

$$T(F) = S_1(f) \cup \ldots \cup S_m(f) \tag{1}$$

$$S_i(f) = \{A_i x + a_i : x \in f\} \tag{2}$$

$$A_i = a\ rotary\ matrix\ \times \begin{bmatrix} \lambda_i l & 0 \\ 0 & \lambda_i d \end{bmatrix} \tag{3}$$

in which $0 < \lambda_i l, \ldots, \lambda_i d < 1$. Here $\lambda_i l$ is a development factor of number i part of the whole, along the l direction. The change defined by Eqs. (2) and (3) is a self-affine change.

We have found that the concept of a self-similar system can also be used to describe the acupuncture system. This will happen if $d = 1$ in the above equations.

There is a close relationship between the acupuncture holographic units and fetus development. The fractal theory provides a tool to describe the two systems.

References

1. Y. Huang, Fractal dimensions of acupoints distribution on the human body surface and fractal structure of acupoint system function, in Fractals and Beyond, M. M. Novak, ed., World Scientific Publishing, Singapore 1998
2. Y. Zhang, The bio-holographic diagnosis and treatment method, Shandong University Publishing, Jinan China, 1988
3. H. Li, Fractal theory and its application in molecular science, Science Publishing, China, 1993

THE FRACTAL PROPERTIES OF THE LARGE-SCALE MAGNETIC FIELDS ON THE SUN

I. I.SALAKHUTDINOVA

Institute of Solar-Terrestrial Physics, P.O.Box 4026,
Irkutsk 33, 664033, Russia, isalakh@iszf.irk.ru

It is a common knowledge that both the solar plasma and the solar magnetic field have a cellular, discrete structure, which is indicative of their possible fractal origin. The questions of fine structuring of the solar magnetoplasma were considered [1], plausible mechanisms of its genesis were suggested, and possibilities of describing its fractal properties were put forward. The existence of a whole hierarchy of sizes of active features in the solar atmosphere has long been established. For example, it is well known that in addition to the granulation (1"-1.5"), such structures as the mesogranulation (8-15"), the supergranulation (30"-40"), giant cells (20°-40°), and supergiant cells (80°-110°) exist on the Sun. The last two features refer to the so-called large-scale organization of solar magnetic fields [2].

The objective of this paper is to investigate the fractal properties of large-scale magnetic fields on different spatial scales.

This study is based on using:

1. Synoptic maps of large-scale magnetic fields compiled by Stenflo [3] from maps of small-scale magnetic fields using the data from the Mt. Wilson observatory (17" resolution) from 1959 to 1976 and from Kitt Peak (1" resolution) from 1959 to 1976. The longitudinal-latitudinal grid occupies in the longitude λ the range from 0° to 360° with steps $\Delta\lambda = 10°$ and in the latitude φ from $-90°$ to 90° with 30 uniform areas by the sine of latitude φ.

2. Synoptic maps of large-scale magnetic fields obtained [4] at the Wilcox observatory (180" resolution) from 1976 to 1998. The longitudinal-latitudinal grid occupies in the longitude λ the range from 0° to 360° with steps $\Delta\lambda = 5°$ and in the latitude φ from $-75°$ to 75° with 30 uniform areas by the sine of latitude φ.

These longitudinal-latitudinal grids were used in calculations in this paper. The method of scaling the variance of temporal and spatial series was used to calculate the fractal dimension of the maps [5]. The method involves calculating the variance $V(s)$ of a given quantity in each cell with area s. As a result of plotting (on $\log-\log$ coordinates) the dependence of the variance of the function $V(s)$ on the cell size s areas with a linear approximation are identified, where possible, as $V(s) = s^a$ and the coefficients a of this approximation are determined by the method of least squares. The coefficients a are associated with the Hurst exponent $H = a/2$ as they were obtained by testing of model temporal and spatial series with known H. The fractal dimension of magnetic field distribution on given two-dimensional map is $D = 3 - H$, where 3 represents the dimensions of space.

The scaling of the variance of magnetic field strengths using Stenflo's maps and the maps from the Wilcox observatory was investigated by this method

304

$(s=\Delta\varphi x\Delta\lambda)$. The study has shown that the spatial distribution of large-scale magnetic fields on the Sun reveals a fractal character. This manifests itself in the power scaling of statistical characteristics of magnetic fields such as the variance. Also, two typical ranges of spatial scales are distinguished, which appear to characterize two systems of solar large-scale magnetic fields organized into giant $(\sim \Delta\varphi x\Delta\lambda < (25°\pm5°)x\ (45°\pm5°))$ and supergiant cells $(\sim \Delta\varphi x\Delta\lambda > (25°\pm5°)x\ (45°\pm5°))$. The fractal dimensions of the space of the strength function of magnetic fields show a high degree of their irregularity and alternation on separate synoptic maps. Time variations of the exponents corresponding to the fractal dimension show a cyclic character, which testifies to a change of statistical and associated physical properties of solar magnetoplasma with cycle phases.

References

1. Cadavid, A.C., Lawrence, J.K., Ruzmaikin, A.A., and Kayleng-Knight, A.: 1994, *Astrophys. J.*, **429**, 392.
2. Ivanov, S.S.: 1993, *Doklady Akademii Nauk*, **332**, 89.
3. Stenflo, J.O.: 1972, *Solar Phys.*, **23, No. 2**, 307.
4. 1998–1999, In: web-site of Stanford: http://quake.stanford.edu/~wso/synoptic.html.
5. Ivanov, V.E.: 1995, *Izvestiya Akademii Nauk. Seriya fizicheskaya*, **59, No. 7**, 29.

FRACTAL ANALYSIS OF TIDE GAUGE DATA

N.K. INDIRA

CSIR Centre for Mathematical Modelling and Computer Simulation, NAL Belur Campus Bangalore - 560 037, India

One of the important consequences of Global warming is the general rise in mean sea level. There have been several recent reviews of sea level rise and the consequence of climate change. Long term changes of mean sea level are called secular changes. The global changes in the mean sea level are called the eustatic changes, whereas the vertical movements of the land are called eperiogenic movements. Out of these causes of sea level rise, only the eustatic rise is universal one. It also appears that tide gauge records contain long period fluctuations (5-100 years) which indicate that the accurate extrapolation of small sea level rise values from the data is very difficult. For nearly a century, relative mean sea level has maintained a steady rise at many tide gauge stations around the world. The relative mean sea level change at a particular location is the difference between eustatic change and any local change in land elevation [1,2].

Weather and climate system has low dimensional attractors. A consistent feature of weather and climate data is that they are aperiodic and their deviations from periodicity cannot be explained by conventional linear models of time series analysis.

It is therefore essential that the nature of the sea level change is analysed and modelled in order to assess the changes from one coastline to other or from the changes taking place in one ocean from the other neighbouring oceans. In this study the approach to analyse and visualize such a comparison is done through calculating the fractal dimension of the time series on annual averages of mean sea level variations for different stations in four countries representing two different oceans regionally.

The tide gauge data on the coastal stations of the whole world is maintained by the Permanent Service of Mean Sea Level (PSMSL) at their site. The annual data on the tide gauge data were obtained through ftp for analysis. Five coastal stations in India namely, Bombay, Cochin, Madras, Vishakhapatnam and Sagar were considered [4] to estimate the fractal dimension of the tide gauge data. Using the same technique for interpolating the data points between given observed points, namely, IFS technique, the fractal dimensions were calculated for other coastal stations in India for which longer time series were available. This was done as an additional exercise to check the variation between the values in the fractal dimensions for the coasts of India. For the construction of the fractal interpolation algorithm, consider a set of data $\{(x_i, z_i) ; i = 0,....,N\}$ and construct an IFS in $\Re^2$ such that its attractor, which is a graph of continuous function $f : [x_0, x_N] \rightarrow \Re$, interpolates the data. The IFS technique is described in detail elsewhere [3]. Consider the IFS of the form $\{\Re^2, \omega_n, n = 1, ..., N\}$, where the maps are affine transformation

$$\omega_n \begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} a_n 0 \\ c_n d_n \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} + \begin{pmatrix} e_n \\ f_n \end{pmatrix}$$

This is constrained by the data

306

$$\omega_n \begin{pmatrix} x_0 \\ z_0 \end{pmatrix} = \begin{pmatrix} x_{n-1} \\ z_{n-1} \end{pmatrix} \text{ and } \omega_n \begin{pmatrix} x_N \\ z_N \end{pmatrix} = \begin{pmatrix} x_n \\ z_n \end{pmatrix}$$

we get

$a_n = (x_n - x_{n-1})/(x_N - x_0),$

$e_n = x_N x_{n-1} - x_0 x_n)/(x_N - x_0),$

$c_n = (z_n - z_{n-1})/(x_N - x_0) - d_n (z_N - z_0)/(x_N - x_0),$

$f_n = (x_N z_{n-1} - x_0 z_n)/(x_N - x_0) - d_n (x_N z_0 - x_0 z_N)/(x_N - x_0)$

where d_n is any real number. If $\sum_{n=1}^{N} |d_n| > 1$, the fractal dimension is the unique real solution D of

$$\sum_{n=1}^{N} |d_n| a_n^{D-1} = 1.$$

It was found that the fractal dimensions when calculated for Kandla, Calcutta (Garden Reach), Diamond Harbour and Tribeni fall in the range of 1.2 and 1.3 as obtained earlier [4]. It can be said that the fractal dimensions obtained for sea level variation in the coasts of India fall in the range between 1.2 and 1.3.

The natural question now is whether such a behaviour is seen in other nearby oceans also, or if there is any difference in the behaviour of the Ocean which may be reflected on the value of the fractal dimension. In the immediate neighbourhood, the tide gauge data on Singapore coastal line were analysed. The three stations, Jurong, Sembawang and Sultan Shaol were considered for analysis and which showed similar range of fractal dimensions. The coast of Thailand where the three stations, Fort Ph. Chimkalao, Ko Lak and Ko Sichang show the fractal dimensions varying between 1.2 and 1.3, which is similar to the results obtained earlier.

To check and compare the fluctuations of sea level with respect to their fractal dimensions, from one ocean to other, the coastal stations of China representing Pacific Ocean were considered. The four stations in China namely Kanmen, Quihuan, Xiamen and Zhapo when analysed through IFS technique give exactly similar results as obtained earlier.

The fractal dimensions of sea level data for both Indian and Pacific Oceans are obtained and found to lie between 1.2 and 1.3. These non-integer values of the dimensions suggest small number of variables representing the system.

References

1. Houghton, J.T. et al, ed., Climate change - The IPCC scientific assessment, Cambridge Univ. Press, Cambridge, 1990
2. Pugh, D. T., Tides, surges and mean sea level, J. Wiley and Sons, New York, 1987
3. Barnsley, M., Fractals Everywhere, Academic Press Inc., San Diego, 1988
4. Indira, N.K., Singh, R.N. and Yajnik, K.S., Fractal analysis of sea level variations in coastal regions of India, *Current Science*, Vol. **70**, No. 8, 719, 1996

Author Index